









371  
7.9.70















# SANKHYĀ : THE INDIAN JOURNAL OF STATISTICS

CONTENTS OF VOLUME SEVENTEEN 1956-57

(All rights reserved)

## PAPERS

	Page
Some contributions to the design of sample surveys. <i>By Tosio Kitagawa</i> ...	1
Quadratic forms in normally distributed random variables. <i>By John Gurland</i> ...	37
A method for discrimination in time series analysis II. <i>By A Rudra</i> ...	51
On the testing of outlying observations. <i>By A. Kudo</i> ...	67
On some quick decision methods in multivariate and univariate analysis. <i>By J. Roy</i> ...	77
On the method of overlapping maps in sample surveys. <i>By Des Raj</i> ...	89
On the recovery of interblock information in varietal trials. <i>By C. Radhakrishna Rao</i> ...	105
Classification and analysis of linked block designs. <i>By J. Roy and R. G. Laha</i> ...	115
On the dual of a PBIB design and a new class of designs with two replications. <i>By C. S. Ramakrishnan</i> ...	133
Fractional replication in asymmetrical factorial designs and partially balanced arrays. <i>By I. M. Chakravarti</i> ...	143
A general class of quasi-factorial and related designs. <i>By C. Radhakrishna Rao</i> ...	165
Two associate partially balanced designs involving three replications. <i>By J. Roy and R. G. Laha</i> ...	175
Some series of balanced incomplete block designs. <i>By D. A. Sprott</i> ...	185
The concept of asymptotic efficiency. <i>By D. Basu</i> ...	193
A note on the determination of optimum probabilities in sampling without replacement. <i>By Des Raj</i> ...	197



	Page
Almost unbiased estimates of functions of frequencies. <i>By J. B. S. Haldane</i> ...	201
Sufficient statistics in elementary distribution theory. <i>By Robert V. Hogg and Allen T. Craig</i> ...	209
Sufficient statistics and orthogonal parameters. <i>By V. S. Huzurbazar</i> ...	217
A note on the multivariate extension of some theorems related to the uni- variate normal distribution. <i>By D. Basu</i> ...	221
The survey research centre of the central bureau of statistics, Sweden. <i>By Tore Dalenius</i> ...	225
Ragnar Frisch on the concept of national income. <i>By Hannan Ezekiel</i> ...	245
A rejoinder. <i>By Ragnar Frisch</i> ...	246
Food statistics. <i>By P. C. Bansil</i> ...	247
Some observations on input-output analysis. <i>By Oskar Lange</i> ...	305
The use of short-term econometric model for Indian economic policy. <i>By J. Tinbergen</i> ...	337
Approximate distribution of certain linear function of order statistics. <i>By K. C. Seal</i> ...	345
On the unboundedness of infinitely divisible laws. <i>By S. D. Chatterjee and R. P. Pakshirajan</i> ...	349
A note on the orthogonal latin squares. <i>By Nikhilesh Bhattacharya</i> ...	351
A new discrete distribution. <i>Ayodhya Prasad</i> ...	353
An experimental method for obtaining random digits and permutations. <i>By John E. Walsh</i> ...	355
On estimating parametric functions in stratified sampling design. <i>By Des Raj</i> ...	361
A note on variance components in multi-stage sampling with varying probabi- lities. <i>By J. Roy</i> ...	367
A note on two stage sampling. <i>By R. Rangarajan</i> ...	373
Method of matching used for the estimation of test reliability <i>By P. K. Bose and S. B. Chaudhuri</i> ...	377
Recommendations for personnel selection in India based on British selection methods in civil service and industry. <i>By Rhea S. Das</i> ...	385



	Page
Isolation of some morale dimensions by factor analysis. <i>By H. C. Ganguli</i> ...	393
Inversion of $25 \times 25$ matrix on a 602A calculating punch <i>By D. Bose and A. Roy</i> ...	401
REPORT	
Indian Statistical Institute : Twentyfourth Annual Report—1955-56	251
CORRIGENDA	
Corrigenda to the paper "Completeness similar regions and unbiased estimation" <i>By E. L. Lehmann and Henry Scheffe</i> ...	250
BOOK REVIEWS	
Survey of fertility and mortality in Poona district : <i>By V. M. Dandekar and K. Dandekar. A. DasGupta</i> ...	99
A study in the analysis of stationary time series : <i>By Herman Wold A. Rudra</i> ...	103
AUTHOR INDEX	
<i>Ayodha Prasad.</i> A new discrete distribution ...	353
<i>Bansil, P. C.</i> Food statistics ...	247
<i>Basu, D.</i> The concept of asymptotic efficiency ...	193
———— A note on the multivariate extension of some theorems related to the univariate normal distribution ...	221
<i>Bhattacharya Nikhilesh.</i> A note on the orthogonal latin square ...	351
<i>Bose, D and Roy, A.</i> Inversion of $25 \times 25$ matrix on 602A calculating punch ...	401
<i>Bose, P. K. and Chaudhuri S. B.</i> Method of matching used for the estimation of test reliability ...	377
<i>Chakravarti, I. M.</i> Fractional replication in asymmetrical factorial designs and partially balanced arrays ...	143
<i>Chatterjee, S. D. and Pakshirajan R. P.</i> On the unboundedness of infinitely divisible laws ...	349
<i>Craig Allen T. and Hogg Robert V.</i> Sufficient theory in elementary distri- bution theory ...	201
<i>Dalenius, Tore.</i> The survey research centre of the central bureau of statistics, Sweden ...	225
<i>Des Raj.</i> On the method of overlapping maps in sample surveys ...	89



	Page
———— A note on the determination of optimum probabilities in sampling without replacement ... ..	197
———— On estimating parametric functions in stratified sampling designs ...	361
<i>Ezekiel, Hannan</i> Ragnar Frisch on the concept of national income ...	245
<i>Frisch, Ragnar</i> A Rejoinder ... ..	246
<i>Ganguli, H. C.</i> Isolation of some morale dimensions by factor analysis ...	393
<i>Gurland, John.</i> Quadratic forms in normally distributed random variables ...	37
<i>Haldane, J. B. S.</i> Almost unbiased estimates of functions of frequencies ...	201
<i>Hogg, Robert V. and Craig Allen T.</i> Sufficient statistics in elementary distribution theory ... ..	209
<i>Huzurbazar, V. S.</i> Sufficient statistics and orthogonal parameters ...	217
<i>Kitagawa, Tosio</i> Some contributions to the design of sample surveys ...	1
<i>Kudo, A.</i> On the testing of outlying observations ... ..	67
<i>Laha, R. G. and Roy, J.</i> Classification and analysis of linked block designs ...	115
———— Two associate partially balanced designs involving three replications	165
<i>Lange, Oskar</i> Some observations on input-output analysis ... ..	305
<i>Pakshirajan R. P. and Chatterjee, S. D.</i> On the unboundedness of infinitely divisible laws ... ..	349
<i>Rao, C. Radhakrishna</i> On the recovery of interblock information in varietal trials ... ..	105
———— A general class of quasifactorial and related designs ... ..	165
<i>Rnagarajan, R.</i> A note on two stage sampling ... ..	373
<i>Roy, A. and Bose, D.</i> Inversion of $25 \times 25$ matrix on a 602A calculating punch	401
<i>Roy, J.</i> On some quick decision methods in multivariate and univariate analysis ... ..	77
———— A note on variance components in multi-stage sampling with varying probabilities ... ..	367
<i>Roy, J and Laha, R. G.</i> Classification and analysis of linked block designs ...	115
———— Two associate partially balanced designs involving three replications	175
<i>Rudra, A.</i> A method for discrimination in time series analysis II ... ..	51
<i>Seal, K. C.</i> Approximate distribution of certain linear function of order statistics ... ..	345
<i>Walsh, John E.</i> An experimental method for obtaining random digits and permutations ... ..	355



# SANKHYĀ

## THE INDIAN JOURNAL OF STATISTICS

*Edited by : P. C. MAHALANOBIS*

---

VOL. 17, PART 1

JUNE

1956

---

### SOME CONTRIBUTIONS TO THE DESIGN OF SAMPLE SURVEYS

*By* TOSIO KITAGAWA

*Kyusyu University, Fukuoka*

*and*

*Indian Statistical Institute, Calcutta*

#### PART IV. EXACT SAMPLING THEORIES AND ANALYSIS OF VARIANCE SCHEMES ASSOCIATED WITH DESIGNS OF SAMPLE SURVEYS\*

##### 1. INTRODUCTION

The current theories are concerned at least as their first approximate approaches with the errors of estimates arising solely from variations of random sampling. In actual large-scale surveys, where various sources of errors should be expected, our theoretical formulations should be so broad as to cover some of the important features such as failures to cover some of the units in the chosen sample, errors of measurements and biases due to the observers. It would not be sound to consider all these features at once; we should rather gradually proceed from the simplest to the most complicated one. Thus in this part, we propose to discuss a certain mathematical formulation in which some sort of statistical inference theories concerning sample surveys could be established. It may be readily observed that our assumptions take into consideration errors of measurements on a unit and also some slight time-changes in the populations which may always be expected in actual situations. To make our analysis simpler we shall begin with the normality assumption, which, however, can be replaced by other more general assumptions. Our main point is to overcome certain characteristic difficulties associated with the finiteness of the population and sampling without replacement by formulating our finite population in a more realistic sense. Among several authors Cochran (1953) suggested the validities of confidence intervals

---

\* Parts I to III published in *Sankhyā*, 14, 317-362.



based on  $t$ -distributions and made some comments in favour of validity of the normal approximation. It should, however, be pointed out that the results due to several authors about finite population cannot be applied to demonstrate any validity of the  $t$ -distribution so far as they are concerned with the case where the size of the sample becomes infinity. In a previous paper (Kitagawa, 1950b) we have discussed an application of the two-sample theory to statistical inferences for finite populations. The standpoint of the two-sample theory may sometimes be useful in this formulation but there still remain some sort of artificialities. Another formulation which seems in some respects more natural is to appeal to a subsampling scheme in which we shall assume an infinite grand population  $\Pi$  from which our finite population  $N$  should be drawn and consequently our sample of size  $n$  should be recognised as a subsample from  $\Pi$ . In fact D. Basu in a seminar held at the Indian Statistical Institute in June 1953 pointed out our assumptions (1<sup>0</sup>), (2<sup>0</sup>) and (3<sup>0</sup>) were just equivalent to this subsampling procedure from the normal grand population  $\Pi$ . One might not yet be perfectly satisfied with the assumption of the existence of one grand population. In fact so far as sample surveys are concerned, there are real difficulties in imagining possible infinite villages, districts and areas in crops. The new assumptions which will be introduced in §2 and from which our inference theories may be developed may form the basis of an approach in which both the characteristic features are taken into consideration, that is, the errors of measurements and the finiteness of the population.

The second aim of this part is to provide the analysis of variance schemes applicable in sample surveys which can be duly discussed only after establishing some sort of exact sampling theories, so far as their applications are treated from the stochastic standpoint.

## 2. SUBSAMPLE FORMULATIONS AND VALIDITIES OF $t$ AND $z$ DISTRIBUTIONS FOR A FINITE POPULATION

Let us introduce the following assumptions:

*Assumption I:* Let us consider a set of  $N$  grand populations

$$\{\Pi_h\} \quad (h = 1, 2, \dots, N)$$

*Assumption II:* From each of the  $N$  grand populations a sample of size one shall be drawn independently, which we shall denote by  $y_h$  ( $h = 1, 2, \dots, N$ ).

*Assumption III:* Let us draw a sample of size  $n$  without replacement from the set of size  $N$ :  $\{y_h\}$  ( $h = 1, 2, \dots, N$ ), and let us denote this sample by  $\{y_i\}$  ( $i = 1, 2, \dots, n$ ).

*Assumption IV:* The grand population  $\Pi_j$  has normal distribution  $N(\xi_j, \sigma^2)$  ( $j = 1, 2, \dots, N$ ) where  $\sigma^2$  is common to all the  $N$  grand populations.



# SOME CONTRIBUTIONS TO THE DESIGN OF SAMPLE SURVEYS

Lemma 4.1: *The joint probability distribution of  $(\dot{y}_1, \dot{y}_2, \dots, \dot{y}_n)$  is for any set of real numbers  $(x_1, x_2, \dots, x_n)$  given by*

$$\begin{aligned} & Pr.\{\dot{y}_1 < x_1, \dot{y}_2 < x_2, \dots, \dot{y}_n < x_n\} \\ &= \frac{1}{{}_N P_n} \sum_{(i_1, i_2, \dots, i_n)} Pr.\{y_{i_1} < x_1\} Pr.\{y_{i_2} < x_2\} \dots Pr.\{y_{i_n} < x_n\} \end{aligned} \quad \dots \quad (2.01)$$

where  $(i_1, i_2, \dots, i_n)$  means a permutation taking  $n$  elements from  $(1, 2, \dots, N)$  and the summation runs through all the permutations  ${}_N P_n$ .

(a) *The joint distribution of the sample mean and sample variance.* We shall give here the joint distribution of the sample mean  $\bar{y} = (\dot{y}_1 + \dot{y}_2 + \dots + \dot{y}_n)/n$  and sample variance  $s^2 = \Sigma(\dot{y}_i - \bar{y})^2/(n-1)$ .

In view of Weibull (1950) and Lemma 4.1, we shall observe

Lemma 4.2: *The characteristic function of the joint distribution of  $\bar{y}$  and  $s^2$  is given by*

$$\frac{1}{{}_N P_n} \sum_{\alpha} f_{\alpha}(t_1, t_2) \quad \dots \quad (2.02)$$

where  $\alpha = (i_1, i_2, \dots, i_n)$  runs through all permutations of  $n$  elements from  $(1, 2, \dots, N)$  and the characteristic function  $f_{\alpha}(t_1, t_2)$  is defined for each fixed permutation  $\alpha$  such that

$$\begin{aligned} f_{\alpha}(t_1, t_2) = & \left(1 - \frac{2\sigma^2 i t_2}{n-1}\right)^{-\frac{n-1}{2}} \exp \left\{ \frac{\lambda_{\alpha}^2 i t_2}{1 - 2\sigma^2 i t_2 (n-1)^{-1}} \right\} \times \\ & \times \exp \{ \bar{\xi}_{\alpha} i t_1 - \sigma^2 t_1^2 (2n)^{-1} \}, \end{aligned} \quad \dots \quad (2.03)$$

where

$$\lambda_{\alpha}^2 = \sum_{j=1}^n (\xi_{i_j} - \bar{\xi}_{\alpha})^2 \quad \dots \quad (2.04)$$

and

$$\bar{\xi}_{\alpha} = n^{-1} \sum_{j=1}^n \xi_{i_j}. \quad \dots \quad (2.05)$$

Our result seems at first glance to be very complicated, but it may be readily observed that the average shown in (2.02) will sometimes make our formulae simplified.

(b) *The t-distribution.* Now let us define the statistic  $t$  by

$$t = \sqrt{n}(\bar{y} - \mu)/s \quad \dots \quad (2.06)$$

and let us find out the distribution function of  $t$ . For each fixed permutation  $\alpha = (i_1, i_2, \dots, i_n)$  we may write

$$t_{\alpha} = \sqrt{n}(\bar{y}_{\alpha} - \mu)/s_{\alpha}. \quad \dots \quad (2.07)$$



Lemma 4.3: For each fixed  $\alpha$ , the probability density function of the joint distribution of  $\bar{y}_\alpha$  and  $s_\alpha$  is given by

$$\left(\frac{n}{2\pi\sigma^2}\right)^{\frac{1}{2}} \exp\left\{-n\frac{(\bar{y}_\alpha - \bar{\xi}_\alpha)^2}{2\sigma^2}\right\} \exp\left\{-\frac{(n-1)\lambda_\alpha^2}{2\sigma^2}\right\} \times \\ \times \sum_{r=0}^{\infty} \frac{1}{r!} \left(\frac{(n-1)\lambda_\alpha^2}{2\sigma^2}\right)^r \frac{(s_\alpha^2)^{\frac{n-1}{2}+r-1} \exp\left\{-\frac{(n-1)s_\alpha^2}{2\sigma^2}\right\}}{\Gamma\left(\frac{n-1}{2}+r\right) \left(\frac{2\sigma^2}{n-1}\right)^{\frac{n-1}{2}+r}} \quad \dots \quad (2.08)$$

This is due to Weibull (1950).

Now we shall turn to the distribution of  $t_\alpha$  which wears the character of non-centrality, as we shall show in Lemma 4.4.

Lemma 4.4: For each fixed  $\alpha$  the elementary probability function of the statistic  $t_\alpha^2/(n-1)$  is given by

$$\phi_\alpha(t_\alpha^2/(n-1)) d(t_\alpha^2/(n-1)) = \exp\left\{-\frac{(n-1)\lambda_\alpha^2}{2\sigma^2}\right\} \exp\left\{-\frac{n\delta_\alpha^2}{2\sigma^2}\right\} \times \\ \times \sum_{r=0}^{\infty} \left(\frac{(n-1)\lambda_\alpha^2}{2\sigma^2}\right)^r \frac{1}{\Gamma(r+1)} \left\{ \sum_{k=0}^{\infty} (-1)^k \left(\frac{n\delta_\alpha^2}{2\sigma^2}\right)^{\frac{k}{2}} \frac{1}{\Gamma\left(\frac{k}{2}+1\right)} \times \right. \\ \left. \times \frac{\Gamma\left(\frac{n+k}{2}+r\right)}{\Gamma\left(\frac{n-1}{2}+r\right) \Gamma\left(\frac{k+1}{2}\right)} \frac{(t_\alpha^2(n-1)^{-1})^{\frac{k}{2}}}{\left(1+t_\alpha^2(n-1)^{-1}\right)^{\frac{n+k}{2}+r}} \right\} d\left(t_\alpha^2(n-1)^{-1}\right) \quad \dots \quad (2.09)$$

where

$$\delta_\alpha = \bar{\xi}_\alpha - \bar{\xi} = n^{-1} \sum_{j=1}^n \xi_{ij} - N^{-1} \sum_{h=1}^N \xi_h. \quad \dots \quad (2.10)$$

The proof may be obtained as follows. First let us write  $\bar{y}_\alpha = \mu + \delta_\alpha + t s_\alpha / \sqrt{n}$  and let us make a change of variables  $(\bar{y}_\alpha, s_\alpha)$  into  $(t_\alpha, s_\alpha)$  in (2.08). The integration of (2.08) with respect to  $s_\alpha$  in  $0 < s_\alpha < \infty$  can be done term by term in the expansion of  $\exp\{-n\delta_\alpha^2 t_\alpha^2 / \sigma^2\}$  into power series.

Lemma 4.5: The characteristic function of the statistic  $t = \sqrt{n}(\bar{y} - \bar{\xi})/s$  is given by

$$f(\tau) = \frac{1}{N P_n} \sum_{\alpha=(i_1, \dots, i_n)} E\{\exp(i\tau t_\alpha)\} \quad \dots \quad (2.11)$$

where

$$E\{\exp(i\tau t_\alpha)\} = \int_{-\infty}^{\infty} e^{i\tau t_\alpha} \phi_\alpha\left(\frac{t_\alpha^2}{n-1}\right) \frac{2t_\alpha}{n-1} dt_\alpha \quad \dots \quad (2.12)$$

with  $\phi_\alpha(t_\alpha^2/(n-1))$  enunciated in (2.09).



# SOME CONTRIBUTIONS TO THE DESIGN OF SAMPLE SURVEYS

## 3. EXAMPLES OF EXACT SAMPLING DISTRIBUTIONS

Our methods are those which were adopted in Part VI of our previous paper (Kitagawa, 1951a) and which appeal to the essential assumptions (1<sup>0</sup>), (2<sup>0</sup>) and (3<sup>0</sup>). Here we shall enunciate the following fundamental exact sampling distributions.

In our finite population, inferences should be concerned with the one in which some inference about the finite population with the elements  $\{y_j\} (j = 1, 2, \dots, N)$  should be given in view of  $\{y_{i_j}\} (i = 1, 2, \dots, n)$ . In the present formulation depending upon our assumptions (1<sup>0</sup>) to (4<sup>0</sup>), we can divide  $\{y_j\}$  into two classes of which one consists of  $\{y_{i_j}\} (j = 1, 2, \dots, n)$ , the other being those remaining for each permutation  $\alpha = (i_1, i_2, \dots, i_n)$ .

Thus for each fixed permutation  $\alpha$ , the difference between the mean of the finite population  $\bar{y} = N^{-1}(y_1 + \dots + y_N)$  and the sample mean  $\bar{y} = n^{-1}(y_{i_1} + y_{i_2} + \dots + y_{i_n})$  is independently distributed according to the normal distribution  $N(\bar{y} - \bar{y}_\alpha, \sigma^2(N-n)(Nn)^{-1})$ .

Consequently we shall have immediately, in view of Lemmas 4.3 and 4.4, the following.

Lemma 4.6: (1) For each fixed permutation  $\alpha$ , the probability density function of the joint distribution of  $\bar{y}_\alpha - \bar{y} = z_\alpha$  and  $s_\alpha$  is given by

$$\left(\frac{Nn}{2\pi(N-n)\sigma^2}\right)^{\frac{1}{2}} \exp\left\{-\frac{Nn(z_\alpha - \delta_\alpha)^2}{2(N-n)\sigma^2}\right\} \exp\left\{-\frac{(n-1)\lambda_\alpha^2}{2\sigma^2}\right\} \times \\ \times \sum_{r=0}^{\infty} \frac{1}{\Gamma(r+1)} \left(\frac{(n-1)\lambda_\alpha^2}{2\sigma^2}\right)^r \frac{(s_\alpha^2)^{\frac{n-1}{2}+r-1} \exp\left\{-\frac{(n-1)s_\alpha^2}{2\sigma^2}\right\}}{\Gamma\left(\frac{n-1}{2}+r\right) \left(\frac{2\sigma^2}{n-1}\right)^{\frac{n-1}{2}+r}}. \quad \dots \quad (3.01)$$

(2) For each fixed permutation  $\alpha$ , let us define the statistic  $t_\alpha^*$  by

$$t_\alpha^* = (\bar{y}_\alpha - \bar{y}) / s_\alpha \sqrt{\frac{N-n}{Nn}}. \quad \dots \quad (3.02)$$

Then the probability density function of  $t_\alpha^{*2}/(n-1)$  is given by

$$\phi_\alpha^*\left(\frac{t_\alpha^{*2}}{n-1}\right) = \exp\left\{-\frac{(n-1)\lambda_\alpha^2}{2\sigma^2}\right\} \exp\left\{-\frac{nN\delta_\alpha^2}{2(N-n)\sigma^2}\right\} \times \\ \times \sum_{r=0}^{\infty} \left(\frac{(n-1)\lambda_\alpha^2}{2\sigma^2}\right)^r \cdot \frac{1}{\Gamma(r+1)} \sum_{k=0}^{\infty} (-1)^k \left(\frac{nN\delta_\alpha^2}{2(N-n)\sigma^2}\right)^k \frac{1}{\Gamma\left(\frac{k}{2}+1\right)} \times \\ \times \frac{\Gamma\left(\frac{n+k}{2}+r\right)}{\Gamma\left(\frac{n-1}{2}+r\right) \Gamma\left(\frac{k+1}{2}\right)} \cdot \frac{\left(t_\alpha^2 (n-1)^{-1}\right)^{\frac{k}{2}}}{(1+t_\alpha^2 (n-1)^{-1})^{\frac{n+k}{2}+r}}. \quad \dots \quad (3.03)$$



Consequently, we shall reach the following theorem which will play a fundamental role concerning the confidence interval of the population mean  $\bar{y}$  and which will show how and under what conditions the current uses of the  $t$ -distribution remain valid.

Theorem 4.1: *The characteristic function of the statistic*

$$t^* = (\bar{y} - \tilde{y}) / s \sqrt{\frac{N-n}{Nn}} \quad \dots \quad (3.04)$$

is given by

$$f(\tau) = \frac{1}{N! P_n} \sum_{\alpha=(i_1, \dots, i_n)} E\{\exp(i\tau t_a^*)\} \quad \dots \quad (3.05)$$

$$\text{where} \quad E\{\exp(i\tau t_a^*)\} = \int_{-\infty}^{\infty} e^{i\tau t_a^*} \phi_a^* \left( \frac{t_a^{*2}}{n-1} \right)_{n-1} \frac{2t_a^*}{n-1} dt_a \quad \dots \quad (3.06)$$

with  $\phi_a^*$  defined in (3.03).

Next let us proceed to make inferences about the variance of a finite population by means of the sample variance. For each assigned permutation  $\alpha = (i_1, i_2, \dots, i_n)$ , let us denote for the sake of simplicity  $\{y_{ij}\} (j = 1, 2, \dots, n)$  by  $y_{11}^{(\alpha)}, y_{12}^{(\alpha)}, \dots, y_{1n}^{(\alpha)}$  and the remaining ones by  $y_{21}^{(\alpha)}, y_{22}^{(\alpha)}, \dots, y_{2, N-n}^{(\alpha)}$ . Let us put

$$ns_1^{(\alpha)^2} = \sum_{k=1}^n (y_{1k}^{(\alpha)} - \bar{y}_{1.}^{(\alpha)})^2 = \chi_1^{(\alpha)^2}, \quad \dots \quad (3.07)$$

$$(N-n)s_2^{(\alpha)^2} = \sum_{k=1}^{N-n} (y_{2k}^{(\alpha)} - \bar{y}_{2.}^{(\alpha)})^2 = \chi_2^{(\alpha)^2}, \quad \dots \quad (3.08)$$

$$s_3^{(\alpha)^2} = (N-n)nN^{-1}(\bar{y}_{1.}^{(\alpha)} - \bar{y}_{2.}^{(\alpha)})^2 = \chi_3^{(\alpha)^2}, \quad \dots \quad (3.09)$$

$$Ns^{(\alpha)^2} = ns_1^{(\alpha)^2} + (N-n)s_2^{(\alpha)^2} + s_3^{(\alpha)^2}. \quad \dots \quad (3.10)$$

Then what we have to infer from  $s_1^{(\alpha)^2}$  is concerned with  $s^{(\alpha)^2}$ . These  $ns_1^{(\alpha)^2}$ ,  $(N-n)s_2^{(\alpha)^2}$  and  $s_3^{(\alpha)^2}$  are known to be independently distributed according to the non-central chi-square distributions whose characteristic functions are

$$(1-2\sigma^2 it)^{-\frac{n-1}{2}} \exp \{S_1^{(\alpha)} it / (1-2\sigma^2 it)\}, \quad \dots \quad (3.11)$$

$$(1-2\sigma^2 it)^{-\frac{N-n-1}{2}} \exp \{S_2^{(\alpha)} it / (1-2\sigma^2 it)\}, \quad \dots \quad (3.12)$$

$$(1-2\sigma^2 it)^{-\frac{1}{2}} \exp \{S_3^{(\alpha)} it / (1-2\sigma^2 it)\} \quad \dots \quad (3.13)$$

where 
$$S_1^{(\alpha)} = \sum_{j=1}^n (m_{1j}^{(\alpha)} - \bar{m}_{1.}^{(\alpha)})^2, \quad \dots \quad (3.14)$$

$$S_2^{(\alpha)} = \sum_{j=1}^{N-n} (m_{2j}^{(\alpha)} - \bar{m}_{2.}^{(\alpha)})^2, \quad \dots \quad (3.15)$$

$$S_3^{(\alpha)} = n(\bar{m}_{1.}^{(\alpha)} - m)^2 + (N-n)(\bar{m}_{2.}^{(\alpha)} - m)^2. \quad \dots \quad (3.16)$$

For our present purpose let us first notice the results (5.24) due to Weibull (1950) whose transformation will yield

**Theorem 4.2:** *The probability density function of the statistic*

$$w = \frac{(n-1) \sum_{i=1}^N (y_i - \bar{y})^2}{(N-1) \sum_{i=1}^n (\dot{y}_i - \bar{y})^2} \quad \dots \quad (3.17)$$

is given by 
$$\frac{1}{N P_n} \sum_{\alpha} g_{\alpha}(w) \quad \dots \quad (3.18)$$

with 
$$g_{\alpha}(w) dw$$

$$= \exp \left\{ -\frac{S_1^{(\alpha)} + S_2^{(\alpha)} + S_3^{(\alpha)}}{2\sigma^2} \right\} \left[ \sum_{r=0}^{\infty} \sum_{s=0}^{\infty} \left( \frac{S_1^{(\alpha)}}{2\sigma^2} \right)^r \left( \frac{S_2^{(\alpha)} + S_3^{(\alpha)}}{2\sigma^2} \right)^s \times \right. \\ \left. \times \frac{1}{r!s!} \cdot \frac{\Gamma\left(\frac{N-1}{2} + r + s\right)}{\Gamma\left(\frac{n-1}{2} + r\right)\Gamma\left(\frac{N-n}{2} + s\right)} \cdot \frac{N-n}{n-1} \cdot \frac{(z(N-n)(n-1)^{-1})^{\frac{N-n}{2} + r - 1}}{(1 + z(N-n)(n-1)^{-1})^{\frac{N-1}{2} + r + s}} \right] dz. \quad \dots \quad (3.19)$$

where 
$$\frac{N-n}{n-1} z = \frac{N-1}{n-1} w - 1 \quad \dots \quad (3.20)$$

and  $z$  runs through  $0 \leq z < \infty$  while  $(n-1)(N-1)^{-1} \leq w < \infty$ .

#### 4. ANALYSIS OF VARIANCE APPLIED TO A FINITE POPULATION

The uses of analysis of variance in designs and analysis of sample surveys are well recognised. To establish some theory of inference we shall introduce the following assumptions:

*Assumption I.* Let us consider a set of  $MN$  grand populations

$$\{\Pi_{rs}\} (r = 1, 2, \dots, M; \quad s = 1, 2, \dots, N).$$



*Assumption II:* From each grand population  $\Pi_{rs}$  a sample of size one shall be independently drawn, which we shall denote by  $y_{rs}$  ( $r = 1, 2, \dots, M$ ;  $s = 1, 2, \dots, N$ ).

*Assumption III:* Let us draw a sample of size  $m$  ( $i_1, i_2, \dots, i_m$ ) from the set  $(1, 2, \dots, M)$  without replacement, and also a sample of size  $n$  ( $j_1, j_2, \dots, j_n$ ) from the set  $(1, 2, \dots, N)$  without replacement. Let these two samplings be independent. Let us consider a sample of size  $mn$   $\{y_{i_h j_l}\}$  ( $h = 1, 2, \dots, m$ ;  $l = 1, 2, \dots, n$ ).

*Assumption IV:* The grand populations  $\Pi_{rs}$  have the normal distributions  $N(\xi_{rs}, \sigma^2)$ , where  $\sigma^2$  is common to all these  $MN$  distributions for  $r = 1, 2, \dots, M$ ;  $s = 1, 2, \dots, N$ .

Thus our subsampling procedure will yield as a set of  $mn$  values which depend upon a combination  $\gamma$  of the two permutations  $\alpha = (i_1, i_2, \dots, i_m)$ , and  $\beta = (j_1, j_2, \dots, j_n)$ .

In order to simplify our notations, let it be assumed that  $i_r = r$  ( $r = 1, 2, \dots, m$ ) and  $j_s = s$  ( $s = 1, 2, \dots, n$ ). Let  $i, j, u$  and  $v$  be natural numbers such that  $1 \leq i \leq M$ ,  $1 \leq j \leq N$ ,  $m+1 \leq u \leq M$  and  $n+1 \leq v \leq N$  respectively. Thereafter a sample  $y_{i_r j_s}$  will be denoted by  $x_{rs}^{(\gamma)}$  while the other  $y$ 's by  $x_{rv}^{(\gamma)}$ ,  $x_{us}^{(\gamma)}$  and  $x_{uv}^{(\gamma)}$  respectively and their respective means by

$$\bar{x}_{rn}^{(\gamma)} \equiv n^{-1} \sum_{s=1}^n x_{rs}^{(\gamma)}, \quad \dots \quad (4.01)$$

$$\bar{x}_{un}^{(\gamma)} \equiv n^{-1} \sum_{s=1}^n x_{us}^{(\gamma)}, \quad \dots \quad (4.02)$$

$$\bar{x}_{r, N-n}^{(\gamma)} \equiv (N-n)^{-1} \sum_{v=n+1}^N x_{rv}^{(\gamma)}, \quad \dots \quad (4.03)$$

$$\bar{x}_{rN}^{(\gamma)} \equiv N^{-1} \sum_{j=1}^N x_{rj}^{(\gamma)}, \quad \dots \quad (4.04)$$

$$\bar{x}_{mN}^{(\gamma)} \equiv N^{-1} \sum_{j=1}^N \bar{x}_{mj}^{(\gamma)} = (mN)^{-1} \sum_{r=1}^m \sum_{j=1}^N x_{rj}^{(\gamma)}, \quad \dots \quad (4.05)$$

$$\bar{x}_{MN}^{(\gamma)} \equiv (MN)^{-1} \sum_{i=1}^M \sum_{j=1}^N x_{ij}^{(\gamma)} = \bar{y}_{..} = \bar{y}_{MN}, \quad \dots \quad (4.06)$$

and similarly for other means such as  $\bar{x}_{ms}^{(\gamma)}$ ,  $\bar{x}_{M-n,s}^{(\gamma)}$ ,  $\bar{x}_{M-m,v}^{(\gamma)}$ ,  $\bar{x}_{Mn}^{(\gamma)}$  and so on.

# SOME CONTRIBUTIONS TO THE DESIGN OF SAMPLE SURVEYS

Now our objects of statistical inferences concerning finite populations will be concerned with all or some of the sums of squares

$$S_T(M, N) \equiv \sum_{i=1}^M \sum_{j=1}^N (y_{ij} - \bar{y}_{MN})^2 \quad \dots \quad (4.07)$$

$$S_R(M, N) \equiv N \sum_{i=1}^M (\bar{y}_{iN} - \bar{y}_{MN})^2 \quad \dots \quad (4.08)$$

$$S_C(M, N) \equiv M \sum_{j=1}^N (\bar{y}_{Mj} - \bar{y}_{MN})^2 \quad \dots \quad (4.09)$$

$$S_W(M, N) \equiv \sum_{i=1}^M \sum_{j=1}^N (\bar{y}_{ij} - \bar{y}_{iN} - \bar{y}_{Mj} + \bar{y}_{MN})^2 \quad \dots \quad (4.10)$$

while our estimators will be all or some of the following sums of squares which can be calculated from our sample

$$S_T^{(\gamma)}(m, n) \equiv \sum_{r=1}^m \sum_{s=1}^n (x_{rs}^{(\gamma)} - \bar{x}_{mn}^{(\gamma)})^2 \quad \dots \quad (4.11)$$

$$S_R^{(\gamma)}(m, n) \equiv n \sum_{r=1}^m (\bar{x}_{rn}^{(\gamma)} - \bar{x}_{mn}^{(\gamma)})^2 \quad \dots \quad (4.12)$$

$$S_C^{(\gamma)}(m, n) \equiv m \sum_{s=1}^n (\bar{x}_{ms}^{(\gamma)} - \bar{x}_{mn}^{(\gamma)})^2 \quad \dots \quad (4.13)$$

$$S_W^{(\gamma)}(m, n) \equiv \sum_{r=1}^m \sum_{s=1}^n (x_{rs}^{(\gamma)} - \bar{x}_{rn}^{(\gamma)} - \bar{x}_{ms}^{(\gamma)} + \bar{x}_{mn}^{(\gamma)})^2 \quad \dots \quad (4.14)$$

respectively.

The relation between  $S_T(M, N)$  and  $S_T^{(\gamma)}(m, n)$  is quite simple. Indeed the division of the totality of  $M \times N$  elements into the four parts which consist of  $m \times n$ ,  $m \times (N - n)$ ,  $(M - m) \times n$  and  $(M - m) \times (N - n)$  elements respectively, will lead to the following analysis of variance:

$$\begin{aligned} S_T(M, N) = & S_T^{(\gamma)}(m, n) + S_T^{(\gamma)}(m, N - n) + \\ & + S_T^{(\gamma)}(M - m, n) + S_T^{(\gamma)}(M - m, N - n) + \\ & + S_B^{(\gamma)}(m, n; M, N), \end{aligned} \quad \dots \quad (4.15)$$

where

$$S_T^{(\gamma)}(m, N - n) = \sum_{r=1}^m \sum_{v=n+1}^N (x_{rv}^{(\gamma)} - \bar{x}_{m, N-n}^{(\gamma)})^2 \quad \dots \quad (4.16)$$

$$S_T^{(\gamma)}(M - m, n) = \sum_{u=m+1}^M \sum_{s=1}^n (x_{us}^{(\gamma)} - \bar{x}_{M-m, n}^{(\gamma)})^2 \quad \dots \quad (4.17)$$



$$S_T^{(\gamma)}(M-m, N-n) = \sum_{u=m+1}^M \sum_{v=n+1}^N (x_{uv}^{(\gamma)} - \bar{x}_{M-m, N-n}^{(\gamma)})^2 \quad \dots \quad (4.18)$$

$$S_B^{(\gamma)}(m, n; M, N) = S_B^{(\gamma)}(m, n) + S_B^{(\gamma)}(M-m, n) + \\ + S_B^{(\gamma)}(m, N-n) + S_B^{(\gamma)}(M-m, N-n) \quad \dots \quad (4.19)$$

in which  $S_B^{(\gamma)}(m, n) \equiv mn(\bar{x}_{mn}^{(\gamma)} - \bar{x}_{MN}^{(\gamma)})^2 \quad \dots \quad (4.20)$

$$S_B^{(\gamma)}(m, N-n) \equiv m(N-n)(\bar{x}_{m, N-n}^{(\gamma)} - \bar{x}_{MN}^{(\gamma)})^2 \quad \dots \quad (4.21)$$

$$S_B^{(\gamma)}(M-m, n) \equiv (M-m)n(\bar{x}_{M-m, n}^{(\gamma)} - \bar{x}_{MN}^{(\gamma)})^2 \quad \dots \quad (4.22)$$

$$S_B^{(\gamma)}(M-m, N-n) \equiv (M-m)(N-n)(\bar{x}_{M-m, N-n}^{(\gamma)} - \bar{x}_{MN}^{(\gamma)})^2. \quad \dots \quad (4.23)$$

As to the relation between  $S_W(M, N)$  and  $S_W(m, n)$ , a similar decomposition may be applied to the quantities  $Z_{ij}^{(\gamma)} = x_{ij}^{(\gamma)} - \bar{x}_{iN}^{(\gamma)} - \bar{x}_{Mj}^{(\gamma)} + \bar{x}_{MN}^{(\gamma)}$  which yield us

$$S_W(M, N) = S_W^{(\gamma)}(m, n) + S_W^{(\gamma)}(m, N-n) + \\ + S_W^{(\gamma)}(M-m, m) + S_W^{(\gamma)}(M-m, N-n) + \\ + S_{WR}^{(\gamma)}(m; n, N-n) + S_{WR}^{(\gamma)}(M-m; n, N-n) + \\ + S_{WC}^{(\gamma)}(n; m, M-m) + S_{WC}^{(\gamma)}(N-n; m, M-m), \quad \dots \quad (4.24)$$

where

$$S_{WR}^{(\gamma)}(m; n, N-n) \equiv \frac{n(N-n)}{N} \sum_{r=1}^M \left( \bar{x}_{rn}^{(\gamma)} - \bar{x}_{mn}^{(\gamma)} - \bar{x}_{r, N-n}^{(\gamma)} + \bar{x}_{m, N-n}^{(\gamma)} \right)^2 \quad \dots \quad (4.25)$$

$$S_{WR}^{(\gamma)}(M-m; n, N-n) \equiv \frac{n(N-n)}{N} \sum_{u=m+1}^M \left( \bar{x}_{un}^{(\gamma)} - \bar{x}_{M-m, n}^{(\gamma)} - \bar{x}_{M-m, N-n}^{(\gamma)} + \bar{x}_{M-m, N-n}^{(\gamma)} \right)^2 \quad \dots \quad (4.26)$$

$$S_{WC}^{(\gamma)}(n; m, M-m) \equiv \frac{m(M-m)}{M} \sum_{s=1}^n \left( \bar{x}_{ms}^{(\gamma)} - \bar{x}_{mn}^{(\gamma)} - \bar{x}_{M-m, s}^{(\gamma)} + \bar{x}_{M-m, n}^{(\gamma)} \right)^2 \quad \dots \quad (4.27)$$

$$S_{WC}^{(\gamma)}(N-n; m, M-m) \equiv \frac{m(M-m)}{M} \sum_{v=n+1}^N \left( \bar{x}_{mv}^{(\gamma)} - \bar{x}_{m, N-n}^{(\gamma)} - \bar{x}_{M-m, v}^{(\gamma)} + \bar{x}_{M-m, N-n}^{(\gamma)} \right)^2 \quad \dots \quad (4.28)$$

Regarding  $S_R(M, N)$  and  $S_R^{(\gamma)}(m, n)$  their relation becomes more complicated so that we cannot separate out  $S_R^{(\gamma)}(m, n)$  as an independent component from  $S_R(M, N)$ . Indeed we can write merely

$$S_R(M, N) = S_R^{(\gamma)}(m; n, N-n) + S_R^{(\gamma)}(M-m; n, N-n) + \\ + S_B^{(\gamma)}(m; n, N-n) + S_B^{(\gamma)}(M-m; n, N-n) \quad \dots \quad (4.29)$$

where

$$S_R^{(\gamma)}(m; n, N-n) \equiv N \sum_{r=1}^m \left( \frac{n(\bar{x}_{rn}^{(\gamma)} - \bar{x}_{mn}^{(\gamma)})}{N} + \frac{(N-n)(\bar{x}_{r, N-n}^{(\gamma)} - \bar{x}_{m, N-n}^{(\gamma)})}{N} \right)^2 \quad \dots (4.30)$$

$$S_R^{(\gamma)}(M-m; n, N-n) \equiv N \sum_{u=m+1}^M \left( \frac{n(\bar{x}_{un}^{(\gamma)} - \bar{x}_{M-m, n}^{(\gamma)})}{N} + \frac{(N-n)(\bar{x}_{u, N-n}^{(\gamma)} - \bar{x}_{M-m, N-n}^{(\gamma)})}{N} \right)^2 \quad \dots (4.31)$$

$$S_B^{(\gamma)}(m; n, N-n) \equiv Nm \left( \frac{n(\bar{x}_{mn}^{(\gamma)} - \bar{x}_{MN}^{(\gamma)})}{N} + \frac{(N-n)(\bar{x}_{m, N-n}^{(\gamma)} - \bar{x}_{MN}^{(\gamma)})}{N} \right)^2 \quad \dots (4.32)$$

$$S_B^{(\gamma)}(M-m; n, N-n) \equiv N(M-m) \left( \frac{n(\bar{x}_{M-m, n}^{(\gamma)} - \bar{x}_{MN}^{(\gamma)})}{N} + \frac{(N-n)(\bar{x}_{M-m, N-n}^{(\gamma)} - \bar{x}_{MN}^{(\gamma)})}{N} \right)^2 \quad (4.33)$$

and similarly for  $S_G(M, N)$ .

Here let it be observed that (1) in each of the right-hand sides of (4.15), (4.24) and (4.29) all the summands are mutually independent among themselves and that (2) each summand is distributed according to the non-central chi-square distribution with its respective non-centrality parameter, where the degrees of freedom are equivalent to those valid in the case of null-hypothesis, while the non-centrality parameters are those corresponding to the finite population formulation and depending upon our permutation. For example the characteristic function of  $S_T^{(\gamma)}(m, n)$  is given by

$$(1-2\sigma^2 it)^{-\frac{mn-1}{2}} \exp \left\{ -\frac{\mu_T^{(\gamma)}(m, n)^2 2\sigma^2 it}{1-2\sigma^2 it} \right\} \quad \dots (4.34)$$

where

$$\mu_T^{(\gamma)}(m, n)^2 \equiv \frac{1}{2\sigma^2} \sum_{r=1}^m \sum_{s=1}^n (\xi_{rs}^{(\gamma)} - \bar{\xi}_{mn}^{(\gamma)})^2. \quad \dots (4.35)$$

The definitions of  $\xi_{rs}^{(\gamma)}$ ,  $\bar{\xi}_{mn}^{(\gamma)}$  are similar to those of  $x_{rs}^{(\gamma)}$ ,  $\bar{x}_{mn}^{(\gamma)}$  derived from  $\{y_{ij}\}$ . Indeed these are defined by operating our permutation on the set  $\{\xi_{ij}\}$ . As to the inferences to  $S_T(M, N)$  by means of  $S_T(m, n)$  and those to  $S_W(M, N)$  by means of  $S_W(m, n)$  there are many common features among them and an ordinary analysis of variance. Our inferences concerning these two cases will depend upon the ratios of two independent non-central chi-squares. Let  $S_1$  and  $S_2$  be two independent non-central chi-squares with the degrees of freedom  $f_1$  and  $f_2$  and with the non-centrality parameters  $\lambda_1$  and  $\lambda_2$  respectively.



Then the probability density function of the distribution of  $S_1/S_2$  will be given by

$$k(z; f_1, \lambda_1; f_2, \lambda_2) = e^{-\frac{\lambda_1 + \lambda_2}{2}} \sum_{r=0}^{\infty} \sum_{s=0}^{\infty} \frac{\lambda_1^r \lambda_2^s}{r! s!} \frac{\Gamma\left(\frac{f_1 + f_2}{2} + r + s\right)}{\Gamma\left(\frac{f_1}{2} + r\right) \Gamma\left(\frac{f_2}{2} + s\right)} \frac{z^{\frac{f_1}{2} + r - 1}}{(1+z)^{\frac{f_1 + f_2}{2} + r + s}} \dots \quad (4.36)$$

We shall here observe immediately

Theorem 4.3: *The characteristic function of the statistic  $S_T(M, N)/S_T(m, n) - 1$  and that of  $S_W(M, N)/S_W(m, n) - 1$  are given by*

$$E \left[ \exp \left\{ i\tau \left( \frac{S_T(M, N)}{S_T(m, n)} - 1 \right) \right\} \right] = \frac{1}{M P_m \times N P_n} \sum_{\gamma} E \left[ \exp \left\{ i\tau \left( \frac{S_T(M, N)}{S_T^{(\gamma)}(m, n)} - 1 \right) \right\} \right], \dots \quad (4.37)$$

$$E \left[ \exp \left\{ i\tau \left( \frac{S_W(M, N)}{S_W(m, n)} - 1 \right) \right\} \right] = \frac{1}{M P_m \times N P_n} \sum_{\gamma} E \left[ \exp \left\{ i\tau \left( \frac{S_W(M, N)}{S_W^{(\gamma)}(m, n)} - 1 \right) \right\} \right], \dots \quad (4.38)$$

where

$$E \left[ \exp \left\{ i\tau \left( \frac{S_T(M, N)}{S_T^{(\gamma)}(m, n)} - 1 \right) \right\} \right] = \int_{-\infty}^{\infty} e^{i\tau z} k \left( z; MN - 1, \mu_T^{(\gamma)}(M, N)^2 - \mu_T^{(\gamma)}(m, n)^2; mn - 1, \mu_T^{(\gamma)}(m, n)^2 \right) dz \dots \quad (4.39)$$

$$E \left[ \exp \left\{ i\tau \left( \frac{S_W(M, N)}{S_W^{(\gamma)}(m, n)} - 1 \right) \right\} \right] = \int_{-\infty}^{\infty} e^{i\tau z} k \left( z; (M-1)(N-1); \mu_W^{(\gamma)}(M, N)^2 - \mu_W^{(\gamma)}(m, n)^2; (m-1)(n-1), \mu_W^{(\gamma)}(m, n)^2 \right) dz \dots \quad (4.40)$$

with

$$\mu_T^{(\gamma)}(M, N)^2 = (2\sigma^2)^{-1} \sum_{i=1}^M \sum_{j=1}^N (\xi_{ij} - \tilde{\xi}_{MN})^2 \quad \dots \quad (4.41)$$

$$\mu_T^{(\gamma)}(m, n)^2 = (2\sigma^2)^{-1} \sum_{r=1}^m \sum_{s=1}^n (\xi_{rs}^{(\gamma)} - \bar{\xi}_{mn}^{(\gamma)})^2, \quad \dots \quad (4.42)$$

$$\mu_W^{(\gamma)}(M, N)^2 = (2\sigma^2)^{-1} \sum_{i=1}^M \sum_{j=1}^N (\xi_{ij} - \tilde{\xi}_{iN} - \tilde{\xi}_{Mj} + \tilde{\xi}_{MN})^2, \quad \dots \quad (4.43)$$

$$\mu_W^{(\gamma)}(m, n)^2 = (2\sigma^2)^{-1} \sum_{r=1}^m \sum_{s=1}^n (\xi_{rs}^{(\gamma)} - \bar{\xi}_{ms}^{(\gamma)} - \bar{\xi}_{rn}^{(\gamma)} + \bar{\xi}_{mn}^{(\gamma)})^2. \quad \dots \quad (4.44)$$

To proceed to  $S_R(M; N)$  we have to prepare the following

**Lemma 4.7:** Let  $\{x_i\}, \{y_i\}$  ( $i = 1, 2, \dots, k$ ) be a set of  $2k$  mutually independent stochastic variables where each  $x_i$  and  $y_i$  are distributed in  $N(a_i, \sigma^2)$  and in  $N(b_i, \sigma^2)$  respectively. Let  $p$  and  $q$  be non-negative real numbers such that  $p^2 + q^2 = 1$ .

Let us now define

$$S_1 \equiv \sum_{i=1}^k (x_i - \bar{x})^2, \quad \dots \quad (4.45)$$

$$S_2 \equiv \sum_{i=1}^k (y_i - \bar{y})^2, \quad \dots \quad (4.46)$$

$$S_3 \equiv \sum_{i=1}^k \left( p(x_i - \bar{x}) + q(y_i - \bar{y}) \right)^2. \quad \dots \quad (4.47)$$

Then the characteristic function of the joint distribution of  $S_1, S_2$ , and  $S_3$  is given by

$$E \left[ \exp \{ i(t_1 S_1 + t_2 S_2 + t_3 S_3) \} \right] = \frac{1}{\Delta_1^{(k-1)/2}} \exp \left\{ -\frac{\Delta}{\Delta_1} \right\}, \quad \dots \quad (4.48)$$

$$\text{where} \quad \Delta_1 \equiv \{1 - 2\sigma^2 i(t_1 + p^2 t_3)\} \{1 - 2\sigma^2 i(t_2 + q^2 t_3)\} + 4p^2 q^2 \sigma^2 t_3^2, \quad \dots \quad (4.49)$$

$$\Delta = \frac{1}{2\sigma^2} \sum_{j=1}^k \begin{vmatrix} 1 - 2\sigma^2 i(t_1 + p^2 t_3) & -2\sigma^2 i p q t_3' & -(a_j - \bar{a}) \\ -2\sigma^2 i p q t_3 & 1 - 2\sigma^2 i(t_2 + q^2 t_3) & -(b_j - \bar{b}) \\ -(a_j - \bar{a}) & -(b_j - \bar{b}) & (a_j - \bar{a})^2 + (b_j - \bar{b})^2 \end{vmatrix} \quad \dots \quad (4.50)$$

where  $\bar{a}$  and  $\bar{b}$  are the arithmetic means of  $\{a_j\}$  and  $\{b_j\}$  respectively.



The proof can be obtained by a direct calculation of the characteristic function. This lemma and the following corollary seem to us to be very important in successive designs of experiments.

Corollary 4.1: *The characteristic function of the joint distribution function of  $S_1$  and  $S_3$  is given by*

$$E[\exp\{i(S_1 t_1 + S_3 t_3)\}] \\ = (1 - 2\sigma^2 i t_1 - 2\sigma^2 i t_3 - 4\sigma^2 q^2 t_1 t_3)^{-\frac{k-1}{2}} \exp \left\{ \frac{2\sigma^2 i \sigma_a^2 t_1 + 2\sigma^2 i t_3 \sum_{j=1}^k (p \delta a_j + q \delta b_j)^2}{1 - 2\sigma^2 i t_1 - 2\sigma^2 i t_3 - 4\sigma^2 q^2 t_1 t_3} \right\} \quad \dots (4.51)$$

where

$$\sigma_a^2 = (2\sigma^2)^{-1} \sum_{j=1}^k (a_j - \bar{a})^2, \quad \dots (4.52)$$

$$\delta a_j = (2^{\frac{1}{2}} \sigma)^{-1} (a_j - \bar{a}), \quad \dots (4.53)$$

$$\delta b_j = (2^{\frac{1}{2}} \sigma)^{-1} (b_j - \bar{b}). \quad \dots (4.54)$$

The essential point to note here is that there are some characteristic features of the exact sampling theory for finite populations which make it necessary to use somewhat coherent statistics as in (4.51).

## 5. SUMMARY

We have established a certain set of exact sampling distributions under certain assumptions which in our opinion will give a more correct picture of real situations. Here we have been content with giving theoretical considerations which in combination with some abbreviated numerical calculations will yield us some useful results applicable to practical cases. Specially, the average taken over all possible permutations will make it clear how and under what conditions ordinary uses of  $t$ - and  $F$ -distributions may be justified. It is also to be noted that our fundamental idea is to appeal to subsample and two-sample formulations, and that, on the contrary, assumptions of normality are rather artificial.

## SOME CONTRIBUTIONS TO THE DESIGN OF SAMPLE SURVEYS

### PART V. OPERATIONAL FORMULATION OF PROBLEMS OF STATISTICAL INFERENCES IN SAMPLE SURVEYS

#### 1. INTRODUCTORY

Different types of errors occur in the case of large-scale sample surveys. Deming (1944, 1950) gave a detailed listing and description of the different types of errors which should be taken into consideration both in designing and analysing sample surveys. Recently Hansen, Hurwitz, Marks, and Mauldin (1951) discussed response errors which are important factors influencing accuracies of surveys. In this Part V, we shall consider the main sources of error in large-scale sample surveys. For this purpose the classification of different types of error into three types made by Mahalanobis (1944, 1946) is specially suited at least for general considerations. Mahalanobis (1944) mentions the following three types of error: (1) sampling fluctuation, (2) observational error and (3) gross inaccuracies, where (1) and (2) may be presumed to follow probabilistic schemes exactly or at least approximately and hence to be amenable to statistical treatment which, however, does not apply to errors of type (3).

"...In actual practice, however, it is difficult to separate these latter two groups, and it is necessary to pool together the second and the third types under one common head which may be called recording mistakes arising from the human factor" (Mahalanobis, 1944). He also "revealed the great importance of controlling and eliminating as far as possible the mistakes which occurred at the stage of the field survey" (Mahalanobis, 1944, p.409). "One way of doing this would be to organise the sample survey in the form of two or more interpenetrating subsamples" (Mahalanobis, 1944, p. 381). In fact, in spite of many controversies over the usefulness of interpenetrating samples, their diagnostic power can only be duly recognised after one has taken into consideration all the three types of errors and not merely (1) and (2). If we consider—and we must consider—situations in which errors of all these three types should be duly treated, our usual formulation will be found to be too narrow and our usual theories of statistical inference will not be sufficient to cover all the problems. How to consider this type of error will be discussed in § 2, whereas § 3 will be devoted to the problem of statistical inference and controls for guarding against this type of error.

#### 2. STATE, OPERATOR AND SCHEME

We now consider a case where the third type of errors viz., gross inaccuracies should be taken into consideration in order that we should be able to give a more realistic formulation of the problem of sampling design. If we confine ourselves to the first type of error, viz. sampling fluctuations, then each element of our population has a definite value which may be a scalar but may sometimes be vectorial. If we consider both the first and the second types of error, then to each element of our



population there should correspond a stochastic variable. so far as the second type of error may be amenable to probabilistic approach.

The third type of error, on the other hand, is actually a very broad type which includes all errors belonging to neither of these two, and naturally involves various kinds of errors; inaccuracies and falsehood in statements and recording, tricks and so on. Thus the sound approach to make a step forward is not to give a too broad (and obscure) formulation aiming to cover all types of errors that could be imagined, but rather to choose, corresponding to each stage of theoretical and practical development, some restricted domain we should take and we could take into consideration to make an adequate and effective improvement of our designs and analysis.

For this purpose, we propose to introduce here the notions of state, operator and scheme.

In almost all sampling surveys, we shall be able to introduce the notions of respondents, investigators and scheme of surveys. To each respondent there corresponds an objective existence which we call a state and which can be recognised to be an existence independent of our surveys. Responses obtained (if possible) from a respondent through some procedure by investigators may result in vector values, because they will give information on each question item of the schedule, giving answers of the nature of either attributes or variables. There will, however, occur problems of non-response and also of possible interference between investigators and respondents. An abstract idea of investigators which may include any other type of questionnaires such as mail or telephone or interview survey should be more relevantly represented by the notion of operators. Thus there is a state  $\xi$  of objective existence and to each state an operator  $\alpha$  will be applied so that it may give us variables under a certain scheme  $S$ . The domain of  $\alpha$  in which, under the scheme  $S$ , we may be able to observe some variable corresponding to a state  $\xi$  does not necessarily cover the whole of possible states. If it be defined for a certain set of state  $\xi$ , operation  $\alpha$  and scheme  $S$ , then we shall denote the variable by  $S(\alpha, \xi)$ .

An abstract idea of  $S(\alpha, \xi)$  will be so broad as to be associated with or to be subject to falsehood, deliberations and even strategies, for which there could not be any objective approach, unless we restrict ourselves to certain realms of  $S$ ,  $\alpha$  and  $\xi$ . Some kinds of falsehood, deliberations and strategies may have naturally various sources. It is impossible to suggest *a priori* a method by which we should be able to measure or to control all sorts of these errors. Nevertheless we think some or all of the following approaches would be particularly useful in dealing with this third type of error.

(a) *Restriction within a certain domain of types of error:* In taking into consideration all sorts of errors at the same time, our attitude should be gradually progressive. At the first stage we may consider the types of error for which the following procedures (b)—(d) are comparatively easier than others.

## SOME CONTRIBUTIONS TO THE DESIGN OF SAMPLE SURVEYS

(b) *Application of randomisation*: The principal object of randomisation is to introduce a probabilistic scheme so that valid statistical inferences can be made. One cannot help appealing to this principle in order to establish any useful results applicable to some class of errors belonging to the third type. Indeed the scheme which Hansen, Hurwitz, Marks and Maulden (1951) used in discussing response errors in surveys and the scheme which Sukhatme and Seth (1952) used in studying non-sampling errors in surveys have this as a common feature that each of them appeals to the principle of randomisation (although the models are different).

(c) *Application of the principle of 'transformation'*: In actual cases it frequently occurs that we are really concerned not with the variables  $A(\alpha, \xi)$  themselves but some differences among these. In these cases we may be able to eliminate certain unmanageable factors, thanks to the operations of differences. For instance, suppose we are concerned with the states at two different times, say,  $A(\alpha, \xi_{t_1})$  and  $A(\alpha, \xi_{t_2})$ . In spite of the fact that for some set of  $A$ ,  $\alpha$  and  $\xi$ ,  $A(\alpha, \xi_{t_1})$  and  $A(\alpha, \xi_{t_2})$  cannot be obtained, it may be possible in such cases to obtain the variables  $A(\alpha, \xi_{t_1} - \xi_{t_2})$  defined for every set of  $A$ ,  $\alpha$  and  $\xi$ .

This principle belongs to the realm of logic and may be regarded as a prototype of analysis of variance, but it also has an intimate relation with the following principle  $\alpha$ . Generally speaking there may sometimes be another function  $\phi$  such that  $A(\alpha, \phi(\xi_{t_1}, \xi_{t_2}))$  can be defined throughout the whole domain or at least in a broader domain. Moreover there are possibilities of making use of a couple or a team of operators by which  $A((\alpha, \beta), \xi)$  may be defined throughout a broader domain. The combination  $(\alpha, \beta)$  means the co-operation of two types of investigators  $\alpha$  and  $\beta$  where  $\alpha$  is a proper investigator who wants to obtain necessary data from respondents while  $\beta$  is an auxiliary person who is not well-trained as an investigator but who has intimate knowledge of the respondents and will serve to make respondents confident enough to answer correctly to  $\alpha$ . Similarly the questionnaire may sometimes contain some set  $B$  of questions which has a similar effect on the respondents as this auxiliary person  $\beta$ . This can be expressed by the symbol  $(A, B)(\alpha, \xi)$ . Instead of starting direct questioning about domestic economies of households, it is often more effective to speak about general topics which lead them naturally to answer the desired questions.

Both before and in course of the sequence of surveys, there sometimes arises the need of some enlightenment and 'education' for respondents by which we can expect to enlarge the domain of  $\alpha$  for which  $A(\alpha, \xi)$  are defined. This domain should be actually denoted by  $A(\alpha, L\xi)$  where  $L$  stands for enlightenment to the respondents. These principles or procedures seem to belong to some sort of expert techniques, but it will not only be possible but also necessary to give theoretical considerations and also to analyse real data. By suitable formulation the efficiencies of such transformations and the costs for executing them should be discussed in a manner similar to the discussion in parts I-III of costs of surveys and variances of estimates where the latter was concerned with the first type of error only.



(d) *Application of operational view-points:* After all our efforts of making use of (a), (b) and (c) there may still remain certain cases in which the domain of  $\xi$  for which  $A(x, \xi)$  is defined for every  $x$  and  $A$  is not coincident with the whole space of  $\xi$ . For example, let us consider a sampling survey on living costs which requires of each sample household to write in their diary the daily expenditures on and the quantities consumed of each item of food in suitable units. All households cannot be expected to agree with writing this sort of diary for several months. Thus, broadly speaking, there arises the problem of non-response so far as we adhere to such uses of diaries. The social and economic circumstances which cause this type of non-response have some relationship with, say, the living standard and we cannot deny that the biases may not be negligible. If our survey should be confined to the diary records. In such a situation our attitude (to be justified from operational point of view) should be to divide the aims of our surveys into two types; the first type will be concerned with the diary reports, while the second type will aim to investigate the circumstances under which some households cannot or will not respond and also to study the relationship of this with their ways of living. So far as it is accompanied by the latter surveys, the first part of the survey is useful for a certain class of operational problems, for example, for wage agreements between the entrepreneur and the labour union. The estimates of living costs obtained from a sample survey of the first part with auxiliary data of the second part may be recognised to have operational value. This was the method adopted by Kitagawa and Fujita (1951) for a sampling survey of living costs of coal miners in 1948.

### 3. OPERATIONAL FORMULATION OF STATISTICAL INFERENCE

In dealing with the third type of errors our general principles illustrated in § 2 should be introduced both in designing the sample survey and in the analysis of data. Regarding the latter it seems necessary (and adequate) to appeal to an operational formulation of the problem of statistical inference. An operational formulation means an elaboration of our statistical decisions so as to deal with some sort of previous knowledge and/or problems of prognosis in a more comprehensive way as we have done in two previous papers (Kitagawa, 1953a, 1953b).

In a previous paper (Kitagawa, 1953a) it was pointed out that certain operational formulations of previous knowledge would sometimes be useful, because of the fact that previous knowledge may be derived from various sources, not necessarily from sampling, or from designed experiments to which probabilistic approaches are possible, but also sometimes from current literature and from obscure sources. Such situations will surely occur when we shall take into consideration the third type of errors discussed in § 1 and § 2. Since there remains usually a lack of objective knowledge enough to appeal to averaging process, however, we may endeavour to apply certain principles enunciated in § 2, and also some statistical procedures which are both operational and objective. Some elaboration of statistical inferences such

## SOME CONTRIBUTIONS TO THE DESIGN OF SAMPLE SURVEYS

as pooling of data (i.e. estimation after preliminary testing of hypothesis) as well as a more general procedure of successive process of statistical inferences discussed in our papers (Kitagawa, 1950a, 1953b) will be found to be specially useful in such situations. In the next section, we shall give an example of applications in connection with interpenetrating samples. Detailed designs of experiments developed in agricultural experimentation may be also helpful in analysing some types of errors in conjunction with psychometric, sociometric and econometric studies on sources of such errors to which sampling survey and design of experiments are now being applied by various research workers.

Successive procedures seem to be extremely useful in dealing with the broad class of errors belonging to the third type. Indeed a follow-up procedure dealing with the non-response problem adopted by Hansen and Hurwitz (1943) and an interesting method of finding sample size due to Birnbaum and Sirken (1950) when non-response is present, appeal to double sampling or to two sample procedure each of which is a special case of successive procedure.

### 4. USES OF INTERPENETRATING SAMPLES

The method of interpenetrating net-works of samples in sample surveys uses a type of design in which the sample units are arranged in two or more independent sets of samples each supplying an independent estimate of the variate under study. This brilliant idea is originally due to Mahalanobis (1940, 1946) and has been recommended by the United Nations Sub-commission on Statistical Sampling. Its purpose is "to provide statistical controls for detecting and guarding against such recording mistakes" (Mahalanobis, 1944) where "recording mistakes" in his terminology comprise the second and the third types of error in §2.

Indeed there have been several authors on both the practical and the theoretical sides who do not recognise this method as a useful statistical control. The present author is unable to discuss the effects of the application of the method, examining in detail the data available. It is, however, intended to show that some of the theoretical objections against the method are not quite appropriate. This is due to not taking the whole role of inter-penetrating samples in consideration fully. Thus although the calculations of the efficiency per unit cost given by Mokashi (1949) shows one of essential loss of efficiency due to the interpenetrating samples, his considerations do not seem to cover the different roles and the functions of these samples. There are various aspects of interpenetrating samples which sometimes lead us to some sort of confusion of notions. To make our views clear in distinction with some authors, let us quote here the description of Ghosh (1949) who discussed the functions of interpenetrating samples in some details and who was not convinced of the actual uses of these samples. According to Ghosh (1949) the basic roles of these samples are enunciated as follows. "Basically the method consists in having two



(or more) samples from the same population so designed that (i) each of these samples will furnish a valid estimate (with its confidence limits) of a common population characteristic; (ii) it will be possible to make statistically valid comparisons between the different samples and (iii) in case the samples (as observed) are not significantly different from one another (as is expected if the samples are enumerated correctly) it will be possible to construct a joint estimate (with confidence limits) of the population characteristic by pooling the information from all the samples". (Ghosh, 1949, pp.108-109). Furthermore he points out (quite appropriately) the following: "It may be noted that the stronger the (positive) correlation between the samples the greater will be the sensitiveness (or discriminating power) of the comparison between the samples, mentioned under (ii) above and the lower will be the efficiency (or precision) of the joint estimate from all the samples, mentioned under (iii)." (Ghosh 1949, p.109). It is thus clear that the uses of interpenetrating samples can be judged not only by the efficiency of the joint estimate but also by the discriminating power of the comparison between the samples, and also that once we enter into the latter, what we are really concerned with is not only the null hypothesis but also the alternative hypothesis, that is, the possibility of different populations must be taken into consideration. Consequently the description of Ghosh seems to us somewhat unsatisfactory.

The discriminating power of interpenetrating samples should be emphasised. The calculation of efficiencies given by Mokashi seems therefore to be inadequate (1949) who takes into consideration the role (iii) merely. "If comparisons between the different investigators by means of interpenetrating samples have been arranged, the comparative results must be available as quickly as possible, in order that effective action may be taken if discrepancies are discovered". (F. Yates, 1949, p.107). Yates also indicates the following important points: "Interpenetrating samples are of value if the survey or census has to be carried out by successive stages. This is frequently necessary when preliminary results are required quickly". (Yates, 1949, p. 44.).

In summing up, our conclusion is that the true merits of interpenetrating samples can only be suitably discussed from the point of view of successive process of statistical inferences and controls in which we must and we shall discuss some elaborated inferences and also effects of statistical controls. In what follows we shall state briefly what we want to mean by this assertion. For the sake of simplicity emphasising the essential points of our views, we shall assume in what follows an infinite normal population, although in real situations we are concerned with finite populations. This simplification is justified by the theory which we have already shown in Part IV as a means of focussing our essential points. The assumption of the normality of our parent population is also a matter of mathematical technique. On the other hand, certain characteristic features of interpenetrating samples should be carefully formulated and introduced in the Assumptions.

# SOME CONTRIBUTIONS TO THE DESIGN OF SAMPLE SURVEYS

## 5. EXAMPLES OF STATISTICAL INFERENCE CONNECTED WITH INTERPENETRATING SAMPLES

Our Assumptions will be as follows:

*Assumption 1.* Let  $(x, y)$  be distributed in a bivariate normal distribution  $\Pi$  with the density function

$$(2\pi\sigma_1\sigma_2)^{-1}(1-\rho^2)^{-\frac{1}{2}} \exp \{-Q/2(1-\rho^2)\}, \quad \dots \quad (5.01)$$

where 
$$Q \equiv \frac{(x-a)^2}{\sigma_1^2} - 2\rho \frac{(x-a)(y-b)}{\sigma_1\sigma_2} + \frac{(y-b)^2}{\sigma_2^2}. \quad \dots \quad (5.02)$$

*Assumption 2.* Let  $O_n : \{(x_i, y_i) \mid i = 1, 2, \dots, n\}$  be a random sample of size  $n$  from the population  $\Pi$ .

*Assumption 3.* The total cost for observing the sample  $O_n$  is equal to  $c_1 f_1(n_1) + c_2 f_2(n_2)$  where  $c_1$  and  $c_2$  are costs per single observation of  $x$  and  $y$  respectively.

The correlation  $\rho$  represents a certain degree of intraclass correlation between the members of a pair,  $x_i$  and  $y_i$ , for the same  $i$ . If we adopt the functional form of the cost of journeys due to Mokashi (1949), then  $f_1(n) = f_2(n) = kn^{\frac{1}{2}}$ ,  $k$  being a certain constant.

In the discussion in §4 the statistical problem connected with interpenetrating samples have been shown to be one of estimation after testing of hypotheses, that is, of the so-called pooling of data. There may be various procedures which resemble real operations actually adopted in practical circumstances. Whatever they may be, the common features are to test statistical hypotheses as a first step. In our formulation we may consider, for example, three different null hypotheses

$$(1^0) H_1 : a = b, \sigma_1^2 = \sigma_2^2; \quad (2^0) H_2 : a = b; \quad (3^0) H_3 : \sigma_1^2 = \sigma_2^2,$$

each of which is composite. The testing of these null hypotheses is the first step. The second step will depend on the results of these. If the tests do not show significant differences the final estimates should be the pooled ones. If, however, differences are significant there may be various procedures possible. For instance, we may consider the following three procedures:

*Procedure A:* If there are certain a priori reasons for preferring the  $x$ -observations to the  $y$ -observations, our procedure will be to adopt the set of observations  $(x_1, x_2, \dots, x_n)$ , the set  $(y_1, y_2, \dots, y_n)$  being completely ignored.

*Procedure B:* There is a possibility of appealing to a successive process of pooling such as was discussed in Part III of the present paper.

*Procedure C:* This process is to draw from the population another random sample of a suitable size which should be regarded with more confidence than either of the samples;  $(x_1, x_2, \dots, x_n)$  or  $(y_1, y_2, \dots, y_n)$ .



For each of these procedures we may give a formulation according to which our statistical procedure will be defined. Parts I and III of the present paper may be of some use after certain generalisations. These generalisations will be in two respects, firstly concerned with the intercorrelation  $\rho$ , while the second one with cost considerations.

Let us now confine ourselves to one special case where  $\sigma_1 = \sigma_2 = \sigma$  (unknown) will be assumed and where the statistical procedure will be of the type (A). This case will arise when there is no difference between the variances of two operators but the presence of some bias of the less trained one which gives us  $y$ -values is expected.

The statistical procedure will be as follows:

(i) Let the statistic  $t$  be defined by

$$t = \frac{\sqrt{n} \bar{d}}{s_d}, \quad \dots (5.03)$$

where we have put

$$\bar{d} = n^{-1} \sum_{i=1}^n d_i = n^{-1} \sum_{i=1}^n (x_i - y_i), \quad \dots (5.04)$$

$$s_d = \{(n-1)^{-1} \sum_{i=1}^n (d_i - \bar{d})^2\}^{\frac{1}{2}}. \quad \dots (5.05)$$

(ii) The estimate  $\bar{x}$  of  $a$  will be defined in the following manner:

(a) If  $|t| < t_{n-1}(\alpha)$ , then

$$\bar{x} = 2^{-1}(\bar{x} + \bar{y}) \quad \dots (5.06)$$

and (b)  $\bar{x} = \bar{x}$ , if otherwise,

$$\dots (5.07)$$

where  $t_{n-1}(\alpha)$  denotes the value of  $t$  with  $n-1$  degrees of freedom for a significance level  $\alpha$ ,  $0 < \alpha < 1$ .

**Theorem 5.1:** *Let  $z$  be any assigned real number. The distribution of  $\bar{x}$  is given by*

$$Pr. \{ \bar{x} < z \} = Pr. \{ \bar{x} < z, |t| < t_{n-1}(\alpha) \} + Pr. \{ \bar{x} < z, |t| \geq t_{n-1}(\alpha) \} \dots (5.08)$$

where we have

$$Pr. \{ \bar{x} < z, |t| < t_{n-1}(\alpha) \} = Pr. \left\{ \frac{\bar{x} + \bar{y}}{2} < z \right\} Pr. \{ |t| < t_{n-1}(\alpha) \} \quad \dots (5.09)$$

with

$$Pr. \left\{ \frac{\bar{x} + \bar{y}}{2} < z \right\} = \frac{(2n)^{\frac{1}{2}}}{(2\pi)^{\frac{1}{2}}\sigma(1+\rho)^{\frac{1}{2}}} \int_{-\infty}^z \exp \left\{ -\frac{n(u-2^{-1}(a+b))^2}{\sigma^2(1+\rho)} \right\} du \quad \dots (5.091)$$

and

$$Pr. \{ |t| < t_{n-1}(\alpha) \} = \frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})\Gamma(\frac{1}{2})} \exp \left\{ -\frac{n(a-b)^2}{2\sigma^2(1-\rho)} \right\} \times \quad \dots (5.092)$$

$$\times \int_0^{t_{n-1}^2(\alpha)/(n-1)} \frac{f^{-\frac{1}{2}}}{(1+f)^{\frac{n}{2}}} {}_1F_1 \left( \frac{n}{2}, \frac{f}{2(1+f)}, \frac{n(a-b)^2}{2\sigma^2(1-\rho)} \right) df, \quad \dots (5.10)$$

$$\begin{aligned} Pr. \{ \bar{x} < z, |t| \geq t_{n-1}(\alpha) \} &= \int_{-\infty}^z \frac{n^{\frac{1}{2}}}{(2\pi)^{\frac{1}{2}}\sigma} \exp \left( -\frac{n(\bar{x}-a)^2}{2\sigma^2} \right) d\bar{x} \times \\ &\times \frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})\Gamma(\frac{1}{2})} \exp \left\{ -\frac{n\{a-b+(1-\rho)(\bar{x}-a)\}^2}{2\sigma^2(1-\rho^2)} \right\} \times \\ &\times 2 \int_{t_{n-1}^2(\alpha)(n-1)^{-1}}^{\infty} \frac{f^{-\frac{1}{2}}}{(1+f)^{\frac{n}{2}}} {}_1F_1 \left( \frac{n}{2}, \frac{f}{2(1+f)}, \frac{n\{(a-b)+(1-\rho)(\bar{x}-a)\}^2}{2\sigma^2(1-\rho^2)} \right) df. \end{aligned}$$

Here  ${}_1F_1$  is the confluent hypergeometric function given by

$${}_1F_1 \left( \frac{n}{2}, \frac{1}{2}, x \right) \equiv \sum_{r=0}^{\infty} \frac{\Gamma(\frac{n}{2} + r) \Gamma(\frac{1}{2})}{\Gamma(r + \frac{1}{2}) \Gamma(r+1)} x^r. \quad \dots (5.11)$$

*Proof:* Let us consider the transformation of  $(x_i, y_i)$  into  $(u_i, v_i)$ :

$$u_i = 2^{-\frac{1}{2}}(x_i - y_i), \quad v_i = 2^{-\frac{1}{2}}(x_i + y_i) \quad \dots (5.12)$$

for  $i = 1, 2, \dots, n$ .

Then we shall have

$$\begin{aligned} q_i &\equiv (2\sigma^2(1-\rho^2))^{-1}((x_i-a)^2 - 2\rho(x_i-a)(y_i-b) + (y_i-b)^2) \\ &= (2\sigma^2(1-\rho))^{-1}(u_i - 2^{-\frac{1}{2}}(a-b))^2 + (2\sigma^2(1+\rho))^{-1}(v_i - 2^{-\frac{1}{2}}(a+b))^2. \quad \dots (5.13) \end{aligned}$$



Consequently the elementary probability associated with the sample of size  $n$  i.e.

$$\prod_{i=1}^n \frac{1}{2\pi\sigma^2(1-\rho^2)^{\frac{1}{2}}} \exp \left\{ -q_i \right\} dx_i dy_i \quad \dots \quad (5.14)$$

will be transformed into

$$\begin{aligned} & \frac{n^{\frac{1}{2}}}{(2\pi)^{\frac{1}{2}} \sigma(1+\rho)^{\frac{1}{2}}} \exp \left\{ -\frac{n(\bar{v}-2^{-\frac{1}{2}}(a+b))^2}{2(1+\rho)\sigma^2} \right\} d\bar{v} \cdot \frac{n^{\frac{1}{2}}}{(2\pi)^{\frac{1}{2}} \sigma(1-\rho)^{\frac{1}{2}}} \exp \left\{ -\frac{n(\bar{u}-2^{-\frac{1}{2}}(a-b))^2}{2(1-\rho)\sigma^2} \right\} d\bar{u} \times \\ & \times G \left( \frac{1}{2(1-\rho)\sigma^2}, \frac{n-1}{2}, S^2 \right) dS^2 \quad \dots \quad (5.14.1) \end{aligned}$$

where we have put

$$\bar{v} = n^{-1} \sum_{i=1}^n v_i, \quad \bar{u} = n^{-1} \sum_{i=1}^n u_i, \quad \dots \quad (5.15)$$

$$S^2 = \sum_{i=1}^n (u_i - \bar{u})^2, \quad \dots \quad (5.16)$$

and the function  $G(\alpha, p, x)$  denotes the gamma distribution

$$G(\alpha, p, x) = \frac{\alpha^p}{\Gamma(p)} e^{-\alpha x} x^{p-1}. \quad \dots \quad (5.17)$$

We now calculate the first and the second term of the right hand side (5.08) separately.

Regarding the first term it will suffice to note that

$$2^{-1}(\bar{x} + \bar{y}) = 2^{-\frac{1}{2}}\bar{v} \quad \dots \quad (5.18)$$

$$t^2 = (n-1)n\bar{u}^2 S^{-2}, \quad \dots \quad (5.19)$$

and to make use of the non-central  $t$ -distribution.

Regarding the second term, we shall first rewrite

$$\begin{aligned} Q &= \frac{n}{2\sigma^2(1-\rho)} \left( \bar{u} - \frac{a-b}{2^{\frac{1}{2}}} \right)^2 + \frac{n}{2\sigma^2(1+\rho)} \left( \bar{v} - \frac{a+b}{2^{\frac{1}{2}}} \right)^2 \\ &= \frac{n}{2\sigma^2} (\bar{x}-a)^2 + \frac{n}{\sigma^2(1-\rho^2)} \left( \bar{u} - \frac{(a-b) + (1-\rho)(\bar{x}-a)}{2^{\frac{1}{2}}} \right)^2 \quad \dots \quad (5.20) \end{aligned}$$

# SOME CONTRIBUTIONS TO THE DESIGN OF SAMPLE SURVEYS

which gives us

$$Pr.\{\bar{x} < z, |t| \geq t_{n-1}(\alpha)\} \quad \dots \quad (5.21)$$

$$= \int \int \int_{D_1} \frac{n^{\frac{1}{2}}}{(2\pi)^{\frac{1}{2}}\sigma} \exp\left\{-\frac{n(\bar{x}-a)^2}{2\sigma^2}\right\} d\bar{x} \cdot G\left(\frac{1}{2(1-\rho^2)\sigma^2}, \frac{n-1}{2}, S^2\right) dS^2 \times \\ \times \frac{(2n)^{\frac{1}{2}}}{(2\pi)^{\frac{1}{2}}\sigma(1-\rho^2)^{\frac{1}{2}}} \exp\left\{-\frac{2n}{2\sigma^2(1-\rho^2)}\left(\bar{u}-\frac{(a-b)+(1-\rho)(\bar{x}-a)}{2^{\frac{1}{2}}}\right)^2\right\} d\bar{u}$$

where the domain of integration is defined by

$$D_1 : \begin{cases} \bar{x} < z, \\ (n-1)n\bar{u}^2S^{-2} \geq t_{n-1}^2(\alpha). \end{cases} \quad \dots \quad (5.22)$$

But the application of the non-central  $t$ -distribution gives us the relation (5.10) where we have introduced  $f = n\bar{U}^2S^{-2}$ ,  $\bar{U}^2$  being equal to  $2^{\frac{1}{2}}u$ .

We can readily observe from Theorem 5.1 the following:

Theorem 5.2: *The  $k$ -th moment of the statistic  $\bar{x}$  is given by*

$$E\{\bar{x}^k\} = (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{\infty} \left\{\frac{a+b}{2} + \left(\frac{1+\rho}{2n}\right)^{\frac{1}{2}}\sigma g\right\}^k \exp\left\{-\frac{g^2}{2}\right\} dg \cdot \{1-H_{n-1}(\alpha, \delta^2(1-\rho)^{-1})\} + \\ + (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{\infty} (a+n^{-\frac{1}{2}}\sigma h)^k \exp\left\{-\frac{h^2}{2}\right\} \cdot H_{n-1}(\alpha^*, (\delta+h)^2(1-\rho^2)^{-1}) dh, \quad \dots \quad (5.23)$$

where we have put

$$H_{n-1}(\alpha, \xi^2) \equiv \frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})\Gamma(\frac{1}{2})} - e^{-\frac{n\xi^2}{2}} \int_{(n-1)^{-1}t_{n-1}^2(\alpha)}^{\infty} \frac{f^{-\frac{1}{2}}}{n} {}_1F_1\left(\frac{n}{2}, \frac{f}{2(1+f)}, \frac{n\xi^2}{2}\right) df \quad \dots \quad (5.24)$$

$$\delta \equiv n^{\frac{1}{2}}(a-b)\sigma^{-1} \quad \dots \quad (4.25)$$

while  $\alpha^*$  is defined by the central  $t$ -distribution with  $n-1$  degrees of freedom such that  $\alpha^*$  is the point whose significance value is equal  $2^{\frac{1}{2}}t_{n-1}(\alpha)$ , that is to say

$$t_{n-1}(\alpha^*) = 2^{\frac{1}{2}}t_{n-1}(\alpha). \quad \dots \quad (5.26)$$



Corrolary 5.1: The mean and the variance of the statistic  $\bar{x}$  are given by

$$E\{\bar{x}\} = \frac{a+b}{2} (1-H_{n-1}) + a\bar{H}_{n-1}^* + \frac{\sigma}{n} \bar{h}H_{n-1}^* \quad \dots (5.27)$$

$$\begin{aligned} \sigma^2\{\bar{x}\} = & \frac{1+\rho}{2n} \sigma^2(1-H_{n-1}) + \frac{\sigma^2}{n} (\bar{h}^2\bar{H}_{n-1}^* - \bar{h}H_{n-1}^*)^2 - \\ & - 2 \frac{\sigma}{n^{\frac{1}{2}}} \left\{ \frac{a+b}{2} (1-H_{n-1}) + a\bar{H}_{n-1}^* \right\} \bar{h}H_{n-1}^* + \left( \frac{a+b}{2} \right)^2 H_{n-1}(1-H_{n-1}) \\ & - 2a \frac{a+b}{2} \bar{H}_{n-1}^*(1-H_{n-1}) + a^2 \bar{H}_{n-1}^*(1-\bar{H}_{n-1}^*). \end{aligned} \quad \dots (5.28)$$

where we have put for the sake of brevity

$$H_{n-1} = H_{n-1}(\alpha, \delta^2(1-\rho)^{-1}) \quad \dots (5.29)$$

$$h^{\nu} \bar{H}_{n-1}^* \equiv (2\pi)^{-1} \int_{-\infty}^{\infty} h^{\nu} \exp \left\{ -\frac{h^2}{2} \right\} H_{n-1} \left( \alpha^*, \frac{(\delta+h)^2}{1-\rho^2} \right) dh, \quad \dots (5.30)$$

for  $\nu = 0, 1, 2$ .

It remains to calculate distribution function of  $\bar{x}$  and its moments numerically. Nevertheless the Theorem and its corollary enable us to make the following observations:

(1) Several authors e.g. Ghosh (1949) points out that the stronger the (positive) correlation  $\rho$  between  $x$  and  $y$  the greater the sensitiveness (or discriminating power) of the comparison between the two samples. Our results bear out these in quantitative terms, for it may be seen from the terms of the non-central  $t$ -distribution that the parameters of non-centrality are  $\delta^2(1-\rho)^{-1}$  and  $(\delta+h)^2/(1-\rho^2)$ , where  $\delta = n^{\frac{1}{2}}(b-a)\sigma^{-1}$ .

(2) Further, several authors e.g. Mokashi (1949), Ghosh (1949) point out also at the same time that the stronger the (positive) correlation  $\rho$  between  $x$  and  $y$  the lower is the efficiency of the joint estimate. Indeed the former showed that the variance of the joint estimate will be  $(1+\rho)$  times that of an estimate based on an independent sample of size  $2n$ . This is not accurate unless we have definitely assigned our statistical procedure. On the contrary, we assert that at least under our formulation the circumstances are not so simple as supposed. One must observe the real situations from (5.28). Indeed there is one term which takes the form  $(1+\rho)\sigma^2k$ , that is, the first term on the right-hand side of (5.28), but it is merely one constituent term of the contributions due to  $\rho$ .

(3) Unless the population means  $a$  and  $b$  are coincident with each other the mere comparison between the variances is not adequate, because the joint estimate

## SOME CONTRIBUTIONS TO THE DESIGN OF SAMPLE SURVEYS

is not an unbiased estimate of  $\alpha$ . It is evident that the mean bias of our estimate  $\bar{x}$  is always less than that of the joint estimate without any preliminary testing of hypotheses.

It is worth while to notice the meaning of the correlation  $\rho$ . Indeed in view of general considerations on effects of biases developed in Cochran (1953, chapter 13) and specially his formulations on interpenetrating subsamples, the formulation due to Hansen, Hurwitz, Marks and Mauldin (1951) and also of our formulation developed in Chapter IV of the present paper, the existence and the magnitude of the correlation  $\rho$  reflect the characteristic features of both finite population and state, operator and scheme formulation, of which more detailed description should be required. We have only pointed out the main features rather in an abstract formulation. The generalised model of analysis of variance discussed in our paper (Kitagawa, 1953) has found its adequacy here.

The correlation  $\rho$  derives sometimes from these generalised models. Thus normal regression theory in the presence of interclass correlation discussed by Halperin (1951) and in a different way from the point of view of successive process of statistical inferences in our paper (Kitagawa, 1953a) will be useful in developing the detailed discussion of interpenetrating samples.

The numerical calculations involved and the detailed description of conditions under which our present formulation will give an approximative picture of real situations will be postponed to another occasion.

## PART VI. THE EFFECTS OF STRATIFICATION

### 1. INTRODUCTORY TO THE PROBLEM OF STRATIFICATION

The problem of constructing a system of strata is different from that of optimum allocation when a certain stratification is given. In this Part we shall consider various problems arising in practice restricting ourselves, however, as a first approach, to non-sequential theories. Our problems will be discussed by introducing a certain mathematical model relating to the objectives and by comparing the effects of various stratifications with reference to these models. It has been a tradition of experts to rely largely on intuition and experience. We shall discuss some of these with the purpose of investigating their merits under the theoretical formulations.

### 2. EFFECTIVE METHODS OF PRELIMINARY STRATIFICATIONS — MAHALANOBIS METHOD

In spite of the elegant theory of optimum allocation in stratified random sampling, there is one serious difficulty, especially at the beginning of a sequence of sample surveys, which may prevent any attempt to adopt the optimum allocation

given in current literature. This is due to the lack of previous knowledge concerning the population which would be necessary for adopting any system of stratification, in particular the knowledge of the within stratum variances. In such a situation there is a method of stratification advocated by Mahalanobis. Broadly speaking, he suggests that when the number of strata is assigned a practical method of possible stratification is to stratify the whole population  $\Pi$  into a set of  $k$  strata  $\{\Pi_i\}$  ( $i = 1, 2, \dots, k$ ) such that the stratum sums are expected to be equal at least approximately, that is, in the notation of Part I

$$N_1\tilde{x}_1 = N_2\tilde{x}_2 = \dots = N_k\tilde{x}_k. \quad \dots (2.01)$$

Since we have no accurate knowledge of the stratum sums, it is clear that we must make use of rough estimates or some values highly correlated with these stratum sums.

One justification of the method may be derived from the elementary fact that under an assigned sum of  $q$ th powers of  $\{x_i\}$ , say  $\sum_{i=1}^k x_i^q = A$ , (1) the minimum value of  $\sum_{i=1}^k x_i^p$  for a certain  $p > q$  and (2) the maximum value of  $\sum_{i=1}^k x_i^p$  for a certain  $p < q$  will be attained when  $x_1 = x_2 = \dots = x_k$  (for each fixed  $k$ ). Thus we easily get

**Lemma 6.1.** *Under the general assumptions in §3 in Part II, let us assume that there exists a real number  $r > 1$  such that*

$$\sum_{j=1}^k (A_j^p c_j)^{\frac{r}{p+1}} = \text{const, say, } M, \quad \dots (2.02)$$

*for all possible stratifications now under consideration.*

*Then among all these systems of stratification the minimum value of the variance with the respective optimum allocations is given by the one which will satisfy*

$$A_1^p c_1 = A_2^p c_2 = \dots = A_k^p c_k, \quad \dots (2.03)$$

*provided that a system of stratification satisfying (2.03) belongs to the class of all systems of stratification considered.*

The stratifications for which the optimum allocation satisfy (2.03) are those for which equal cost will be divided into each stratum in their optimum allocation. The case when  $pr/(1+p) = 1/2$  is specially important because they are concerned with linear combinations of stratum means, and it may be worth noting that there will appear another restriction to Mahalanobis's advocacy as to their cost function.



# SOME CONTRIBUTIONS TO THE DESIGN OF SAMPLE SURVEYS

Indeed we have

Corollary 6.1: In Lemma 6.1, let us assume in particular that  $r = (p+1)/2p$ , that is to say,

$$\sum_{j=1}^k (A_j^p C_j)^{\frac{1}{2p}} = \text{Const.} \quad \dots (2.04)$$

Then our conclusion to Lemma 6.1 holds true, so far as  $0 < p < 1$ .

We are now in a position where either of two courses can be taken into consideration. The first course is to justify the advocacy of Mahalanobis, that is, to give some sufficient conditions in verifying the advocacy which we shall enunciate in Theorem 6.1. The second course is to investigate in more detail a more generalised principle which our mathematical analysis will suggest us and of which Mahalanobis has been surely conscious as seen in Lemma 6.1, that is, the *principle of equipartition* of total cost in each stratum. As to the first course, we may enunciate

Theorem 6.1: Let  $k$  be an assigned positive integer. Let  $S = \{S_r\}$  be a set of stratifications with the  $k$  strata for which the cost function (3.02) in Part II and the condition (2.02) in this Part VI will hold true with a certain power  $p$  such that  $0 < p < 1$ . Let the cost per unit be assumed to be a common value  $c$  which is independent of strata for any  $S_r$  belonging to  $S$ . Let  $S_0$  be a stratification for which  $A_1 = A_2 = \dots = A_k$  holds true.

Then  $S_0$  is a stratification for which the variance of our estimates in its optimum allocation is not greater than any one of  $S_r$  belonging to  $S$ .

Furthermore specially when  $A_i = (N_i \sigma_i)^2$  and when at the same time, it holds true that

$$\frac{\sigma_1}{\tilde{x}_1} = \frac{\sigma_2}{\tilde{x}_2} = \dots = \frac{\sigma_k}{\tilde{x}_k} \quad \dots (2.05)$$

for any stratification  $S_r$  belonging to  $S$ , then  $S_0$  can be characterised by (2.01).

This theorem is of course a formal one for which some detailed discussions should be required, because we have no precise knowledge concerning every population parameter and cost functions in the beginning of our surveys, but it may be still useful to make clear the underlying conditions in justifying Mahalanobis' ideas.

Let us now turn to the second course. In our real situations we have to appeal to a set of approximate values  $\{A'_{ij}\}$  and  $\{c'_{ij}\}$  in order to make stratification so as to satisfy the conditions at least approximately

$$A_1'^p c_1' = A_2'^p c_2' = \dots = A_k'^p c_k' = \left(\frac{M}{k}\right)^{\frac{p+1}{r}} \quad \dots (2.06)$$

Consequently our actual variance to be expected under the adoption of equi-distribution principle should be

$$V_i = \frac{1}{C_p^i} \left( \frac{M}{k} \right)^{\frac{p+1}{pr}} \left( \sum_{j=1}^k \frac{c_j}{c'_j} \right)^{\frac{1}{p}} \left( \sum_{j=1}^k \frac{A_j^{\frac{p+1}{k}}}{A_j^{\frac{1}{(p+1)}}} \right). \quad \dots (2.07)$$

On the other hand when we have complete knowledge about parameters, we have

$$V_p = \frac{1}{C_p^i} \left( \frac{M}{k} \right)^{\frac{p+1}{rp} \frac{p+1}{p}}. \quad \dots (2.08)$$

### 3. PRINCIPLE OF EQUIPARTITION OF TOTAL COST INTO EACH STRATUM

In view of (2.06) and (2.07) we can observe the following assertions which may have some practical interest in themselves.

3.1. *The optimum determination of the number of strata:* Let us consider the case when all  $c_j$  are coincident with a common value  $c_0(k)$  and hence all  $A_j$  with a common value  $A_0(k)$ . Under our assumption of having applied the principle of equipartition of total cost into each stratum we have

$$k A_0(k)^{\frac{pr}{p+1}} c_0(k)^{\frac{r}{p+1}} = M. \quad \dots (3.01)$$

Let us also put

$$k A_0(k)^{\frac{pr}{p+1}} = T, \quad \dots (3.02)$$

and assuming that  $T$  is independent upon of  $k$ , we shall have

$$V_p = \frac{T^{\frac{p+1}{pr}}}{C^{1/p}} \left\{ c_0(k) k^{\frac{r-1}{r}(1+p)} \right\}^{\frac{1}{p}} \quad \dots (3.03)$$

The assumption that  $T$  is independent of  $k$  holds true under the conditions of Theorem 6.1.

Now the problem how the number  $k$  of strata may be determined is seen to be reduced to minimizing

$$\varphi(k) = c_0(k) k^{\frac{r-1}{r}(1+p)} \quad \dots (3.05)$$

where  $k$  runs through the domain  $k \geq 1$ .

## SOME CONTRIBUTIONS TO THE DESIGN OF SAMPLE SURVEYS

The investigation of cost functions will be done in various sampling surveys, and here  $k$  may be considered roughly inversely proportional to the area. Thus as a first approximation we may put here

$$c_0(k) = a + \frac{b}{k^q} \quad \dots \quad (3.06)$$

where  $a$ ,  $b$  and  $q$  are positive constants independent of  $k$ . The value of  $k$  which minimizes (3.06) and hence  $V_p$  is now given by

$$k_0 = \left\{ \frac{b}{a} \left( \frac{qr}{(r-1)(1+p)} - 1 \right) \right\}^{\frac{1}{q}} \quad \dots \quad (3.07)$$

provided that  $q > (r-1)(1+p)r^{-1}$  and  $k_0$  gives us an answer so far as  $k_0 > 1$ .

3.2. *The loss of efficiency by making use of approximate values  $\{A'_j\}$  and  $\{c'_j\}$  in designing sample surveys on the principle of equipartition:* This can be defined by

$$\begin{aligned} e(A', c'; A, c) &\equiv \frac{V_p}{V_i} = \frac{k^{\frac{1+p}{p}}}{\left( \sum_{j=1}^k \frac{c_j}{c'_j} \right)^{\frac{1}{p}} \left( \sum_{j=1}^k \frac{A'_j^{\frac{1}{p+1}}}{A_j^{\frac{1}{p+1}}} \right)} \\ &= \frac{k^{\frac{1+p}{p}}}{\left( \sum_{j=1}^k \frac{A_j^{\frac{p}{p+1}}}{A_j^{\frac{p}{p+1}}} \right)^{\frac{1}{p}} \left( \sum_{j=1}^k \frac{A'_j^{\frac{1}{p+1}}}{A_j^{\frac{1}{p+1}}} \right)} \quad \dots \quad (3.08) \end{aligned}$$

In the case of Mahalanobis principle, we shall have

$$\begin{aligned} e(A', c'; A, c) &= \frac{k^{\frac{1+p}{p}}}{\left( \sum_{j=1}^k \frac{\sigma_j^{2p}}{\sigma_j'^{2p}} \right)^{\frac{1}{p}} \left( \sum_{j=1}^k \frac{\sigma_j'^{\frac{2}{p+1}}}{\sigma_j^{\frac{2}{p+1}}} \right)} \quad \dots \quad (3.09) \end{aligned}$$

and under the assumption that (2.04) should be assumed not only for the population



parameter  $\{\sigma_j\}$  and  $\{\tilde{x}_j\}$  but also for their respective approximate values  $\{\sigma'_j\}$  and  $\{\tilde{x}'_j\}$ , we may write more briefly

$$e(A', c'; A, c) = \frac{k^{\frac{1+p}{p}}}{\left(\sum_{j=1}^k (\tilde{x}_j \tilde{x}'_j)^{-1} \right)^{\frac{1}{p}} \left(\sum_{j=1}^k (\tilde{x}'_j \tilde{x}_j)^{-\frac{2}{p+1}} \right)^{\frac{2}{p+1}}} . \quad \dots (3.10)$$

#### 4. CONSTRUCTION OF POPULATIONS WITH NEARLY CONSTANT COEFFICIENTS OF VARIATION IN DIFFERENT STRATA

Our purpose in this section is to give examples of populations possessing the properties stated in the title.

Let us assume that our whole population consists of an aggregate of some elementary clusters and that any stratum which may be our real concern is also an aggregate of these elementary clusters. Our fundamental idea is to set up elementary clusters whose distributions are all of the Gibrat type, that is, log-normal distributions such as

$$g_{ij}(x) = (2\pi)^{-1}(\sigma x)^{-1} \exp \{-(2\sigma^2)^{-1}(\log x - \xi_{ij})^2\} \quad \dots (4.01)$$

where  $\sigma^2$  is a value common to all elementary clusters.

Let the size of this cluster  $\Pi_{ij}$  be denoted by  $N_{ij}$  and let us define a system of stratification where each stratum  $\Pi_i$  is an aggregate of  $\{\Pi_{ij}\}$  ( $j = 1, 2, \dots, M_i$ ) and by which our total population is now stratified into a sum of  $\{\Pi_i\}$  ( $i = 1, 2, \dots, k$ ). Since the mean and the variance of distribution of (4.01) are equal to

$$E\{X_{ij}\} = e^{\xi_{ij} + \frac{\sigma^2}{2}} \quad \dots (4.02)$$

$$V\{X_{ij}\} = e^{2\xi_{ij} + \sigma^2} (e^{\sigma^2} - 1), \quad \dots (4.03)$$

the following may be readily observed:

(1) The mean of the  $i$ -th stratum is

$$E\{X_{i\cdot}\} = \sum_{j=1}^{M_i} p_{ij} E\{X_{ij}\} = e^{\frac{\sigma^2}{2}} \sum_{j=1}^{M_i} p_{ij} e^{\xi_{ij}} . \quad \dots (4.04)$$

# SOME CONTRIBUTIONS TO THE DESIGN OF SAMPLE SURVEYS

(2) The variance within the  $i$ -th stratum is

$$\begin{aligned} V\{X_{i\cdot}\} &= \sum_{j=1}^{M_i} p_{ij} (E\{X_{ij}\} - E\{X_{i\cdot}\})^2 + \sum_{j=1}^{M_i} p_{ij} V\{X_{ij}\} \\ &= e^{\sigma^2} \left\{ e^{\sigma^2} \sum_{j=1}^{M_i} p_{ij} e^{2\xi_{ij}} - \left( \sum_{j=1}^{M_i} p_{ij} e^{\xi_{ij}} \right)^2 \right\} \quad \dots (4.05) \end{aligned}$$

(3) The square of the coefficient of variation within the  $i$ -th stratum

$$\frac{V\{X_{i\cdot}\}}{E^2\{X_{i\cdot}\}} = \frac{e^{\sigma^2} \sum_{j=1}^{M_i} p_{ij} e^{2\xi_{ij}}}{\left( \sum_{j=1}^{M_i} p_{ij} e^{\xi_{ij}} \right)^2} - 1 \quad \dots (4.06)$$

where we have put for a moment

$$p_{ij} = N_{ij} (N_{i1} + N_{i2} + \dots + N_{iM_i})^{-1} \quad \dots (4.07)$$

for  $j = 1, 2, \dots, M_i; \quad i = 1, 2, \dots, k.$

Now concerning the ratios (4.06) we have the following assertions

(a) We have always

$$e^{\sigma^2} - 1 \leq \frac{V\{X_{i\cdot}\}}{E^2\{X_{i\cdot}\}} \leq e^{\sigma^2} \frac{\max_{1 \leq j \leq M_i} \{e^{2\xi_{ij}}\}}{\left( \sum_{j=1}^{M_i} p_{ij} e^{\xi_{ij}} \right)^2} - 1 \quad \dots (4.08)$$

(b) In particular when  $\xi_{ij}$  can be expressed as  $\xi_{ij} = \mu_i + \nu_j$ , the coefficients of variations are independent of  $\mu_i$  and merely dependent upon  $\nu_j$  and  $p_{ij}$  such as

$$\frac{V\{X_{i\cdot}\}}{E^2\{X_{i\cdot}\}} = \frac{e^{\sigma^2} \sum_{j=1}^{M_i} p_{ij} e^{2\nu_j}}{\left( \sum_{j=1}^{M_i} p_{ij} e^{\nu_j} \right)^2} - 1. \quad \dots (4.09)$$

These assertions (a) and (b) are sufficient to observe under what conditions our stratifications will satisfy at least approximately the condition imposed in 1 of this Part VI. It may be interesting to consider a fine structure of our population by which a set of  $h$  systems of stratification each of which will approximately satisfy the

conditions of constant coefficients of variations will be suggested. Our fine structure may be defined by a decomposition of our whole population  $\Pi$  into the strata

$$\Pi = \sum_{i_1=1}^{k_1} \sum_{i_2=1}^{k_2} \dots \sum_{i_h=1}^{k_h} \Pi_{i_1 i_2 \dots i_h} \quad \dots \quad (4.10)$$

where the sizes of  $\Pi_{i_1 i_2 \dots i_h}$  will be denoted by  $N_{i_1 i_2 \dots i_h}$  and they are assumed to be distributed in Gibrat distributions with the parameters  $\xi_{i_1 i_2 \dots i_h}$  and common  $\sigma^2$ . Our fine structure is now assumed to be as follows:

$$\xi_{i_1 i_2 \dots i_h} = \mu_{1i_1} + \mu_{2i_2} + \dots + \mu_{hi_h} \quad \dots \quad (4.11)$$

for all combinations of  $i_1, i_2, \dots, i_{h-1}$  and  $i_h$ .

Indeed this is nothing but a  $h$ -way classification which will suggest  $h$  systems of stratifications each of which will satisfy at least nearly our conditions.

Now we shall proceed to a multi-dimensional Gibrat distribution because we in actual practice consider other quantities which are highly correlated with our variable and which may be also recognised to be distributed in Gibrat distributions. Indeed in applying the principle of equipartitions and that of Mahalanobis, our reference will be concerned with some other variables highly correlated, for example, with the stratum totals of each possible stratification which can be derived from a previous survey. For the sake of simplicity, let us consider here two-dimensional formulations.

Let us now consider a two-dimensional Gibrat distribution according to which two variables ( $X_{ij}, Y_{ij}$ ) of each elementary cluster  $\Pi_{ij}$  is distributed, namely, the distribution such that  $\log X_{ij}$  and  $\log Y_{ij}$  are distributed in a bivariate normal distribution with their means  $\xi_{ij}$  and  $\eta_{ij}$ , their variances  $\sigma_1^2$  and  $\sigma_2^2$  and their correlation coefficient  $\rho$ , where we shall assume  $\sigma_1^2, \sigma_2^2$ , and  $\rho$  are common to all elementary clusters. Then the covariance between  $X_{ij}$  and  $Y_{ij}$  is given by

$$\sigma(X_{ij}, Y_{ij}) = \sigma_{ij} = e^{\xi_{ij} + \frac{\sigma_1^2}{2} + \eta_{ij} + \frac{\sigma_2^2}{2}} (e^{\rho \sigma_1 \sigma_2} - 1). \quad \dots \quad (4.12)$$

Now what we are to discuss is the effect of mixing upon the correlation coefficient  $\rho(X_{i\cdot}, Y_{i\cdot})$  between  $X_{i\cdot}$  and  $Y_{i\cdot}$  defined for the stratum  $\Pi_i$ . It may be observed that after an amalgamation of these elementary clusters  $\{\Pi_{ij}\}$  into  $\{\Pi_i\}$  correlation coefficient  $\rho_i$  resembles closely that of the elementary cluster, because the latter is equal to

$$\rho(X_{ij}, Y_{ij}) = \frac{e^{\rho \sigma_1 \sigma_2} - 1}{(e^{\sigma_1^2} - 1)^{\frac{1}{2}} (e^{\sigma_2^2} - 1)^{\frac{1}{2}}} \quad \dots \quad (4.13)$$



# SOME CONTRIBUTIONS TO THE DESIGN OF SAMPLE SURVEYS

while as to the latter we have

$$\rho(X_{ij}, Y_{ij}) = \frac{e^{\rho\sigma_1\sigma_2} q_{ij}(\xi, \eta)}{e^{\frac{\sigma_1^2}{2} + \frac{\sigma_2^2}{2}} \{q_{ij}(\xi, \xi)\}^{\frac{1}{2}} \{q_{ij}(\eta, \eta)\}^{\frac{1}{2}}} \times$$

$$\times \frac{1 - q_{ij}(\xi, 1)q_{ij}(\eta, 1)q_{ij}(\xi, \eta)^{-1}}{\left(1 - e^{-\sigma_1^2} \frac{q_{ij}(\xi, 1)^2}{q_{ij}(\xi, \xi)}\right)^{\frac{1}{2}} \left(1 - e^{-\sigma_2^2} \frac{q_{ij}(\eta, 1)^2}{q_{ij}(\eta, \eta)}\right)^{\frac{1}{2}}} \quad \dots \quad (4.14)$$

where we have put

$$q_{ij}(\xi, \eta) = \sum_{j=1}^{M_i} p_{ij} e^{\xi_{ij}} e^{\eta_{ij}}, \quad \dots \quad (4.15)$$

$$q_{ij}(\xi, 1) = \sum_{j=1}^{M_i} p_{ij} e^{\xi_{ij}} \quad \dots \quad (4.16)$$

and similarly for other notations. Similar assertions to those given in (a) and (b) hold true for  $\rho(X_{ij}, Y_{ij})$  and analysis of variance scheme associated with a fine structure will reveal how this value may depend upon  $\xi$ 's and  $\eta$ 's.

## REFERENCES

- BIRNBAUM, Z. W., and SIRKEN, M. G. (1950): Bias due to non-availability in sampling surveys. *J. Amer. Stat. Ass.*, **45**, 48-111.
- COCHRAN, W. G. (1953): *Sampling Techniques*. John Wiley and Sons, New York.
- DEMING, W. E. (1944): *Some Theory of Sampling*. John Wiley and Sons, New York.
- (1950): On errors in surveys. *Amer. Sociological Review*, **9**, 356-369.
- GHOSH, B. (1949): Interpenetrating networks of samples. *Cal. Stat. Ass. Bull.*, **2**, 108-119.
- HALPERIN, M. (1951): Normal regression theory in the presence of interclass correlation. *Ann. Math. Stat.*, **22**, 573-580.
- HANSEN, H. M. and HURWITZ, W. N. (1943): On the theory of sampling from finite populations. *Ann. Math. Stat.*, **14**, 333-362.
- HANSEN, H. M., HURWITZ, W. N., MARKS, E. S. and MAULDIN, W. P. (1951): Response errors in surveys. *J. Amer. Stat. Ass.*, **46**, 147-190.
- KITAGAWA, T. (1950a): Successive process of statistical inferences. (1) *Mem. Fract. Sci.*, Kyushu Univ., Ser. A., **5**, 139-180.
- (1950b): Estimation-formulae used in the sampling survey of the Fishing Catches in Fukuoka Prefecture: Appendix to Sampling survey of fishing catch in Fukuoka Prefecture. Statistics and Research Division, Agricultural Improvement Bureau Ministry of Agriculture and Forestry, 1-39.
- (1951a): Successive process of inferences. (2) *Mem. Frac. Sci.*, kyusyn Univ., Ser A. **6**. 56-95.

- (1953a) : Successive process of statistical inferences. (5) *Mem. Fract. Sci.*, Kyusyu Univ., Ser. A., 7, 95-120.
- (1953b) : Successive process of statistical inferences. (6) *Mem. Fract. Sci.*, Kyusyu Univ., Ser. A., 8,
- KITAGAWA, T. and FUJITA, T. (1951) : *Sampling Surveys of living costs of labourers in Mieke Coal mine* (in Japanese). Toyokeizai Shimpo, Tokyo.
- MAHALANOBIS, P. C. (1940) : A sampling survey of the acreage under jute in Bongal. *Sankhyā*, 4, 511-530.
- (1944) : On large-scale sample surveys. *Phil. Trans. Roy. Soc.*, B231, 329-451.
- (1946) : Recent experiments in statistical sampling in the Indian Statistical Institute. *J. Roy. Stat. Soc.*, 109, 325-370.
- MOKASHI, V. K. (1949) : A note on interpenetrating samples. *J. Indian Soc. Agr. Stat.*, 2, 189-195.
- SUKHATME, P. V. and SETH, G. R. (1952) : Non-sampling errors in surveys. *J. Indian Soc. Agr. Stat.*, 4, 5-41.
- WEIBULL, M. (1950) : The distribution of the  $t$  and  $z$  variables in the case of stratified sample with individuals taken from normal parent populations with varying means. *Skandinavisk Aktuarietidskrift*, Heft 3-4, 137-167.
- YATES, F. (1949) : *Sampling Methods for Censuses and Surveys*, Charles Griffin and Co., London.

*Paper received : August, 1953.*

# QUADRATIC FORMS IN NORMALLY DISTRIBUTED RANDOM VARIABLES

By JOHN GURLAND

*Iowa State College*

## 1. SUMMARY

By applying the inversion formula for characteristic functions convergent series are developed for the distribution of a quadratic form and a ratio of quadratic forms respectively. Bounds are obtained for the errors which result in approximating to the respective distributions by partial sums of the series. Similar methods are applied to the moment-generating function to obtain a convergent series for the expectation of a ratio of quadratic forms and to obtain a bound on the remainder term.

## 2. INTRODUCTION

In two previous papers (Gurland, 1953, 1955) the author has obtained, *inter alia*, convergent Laguerrian expansions for the distribution function of a quadratic form in normally distributed random variables. In neither of these papers is the magnitude of the error term considered which results in taking a partial sum of the series as an approximation to the distribution function. In the present article the methods employed in Gurland (1955) are pursued further to investigate the remainder terms involved in approximating to the distribution of a quadratic form and to the distribution of a ratio of such forms. It will become apparent that if the characteristic roots corresponding to these forms are sufficiently close in a certain sense a bound can be obtained for the remainder term in each case.

In a paper by Pitman and Robbins (1949) the notion of a mixture of distributions is utilized to obtain convergent expansions for the distribution of a positive-definite quadratic form and certain ratios of such forms respectively. The present article differs from this in the method of approach and consequently in the resulting expansions and in the bounds obtained. As in Gurland (1955), the methods employed here utilize and modify a device of Bhattacharya (1945) who considered a special case of the distribution of a quadratic form. Further material developed below pertains to the expectation of a ratio of quadratic forms.

As far as the distribution of a quadratic form in normally distributed random variables is concerned there is no loss of generality in assuming that  $X_1, X_2, \dots, X_n$  are independently and normally distributed with mean zero and variance one and that the quadratic form is  $\sum_{i=1}^n \lambda_i X_i^2$ . In the following section an equivalent but perhaps



more convenient form of the series obtained in Gurland (1955) is developed. This series will converge whatever be the values of  $\lambda_i > 0$ ; however a uniform bound for the remainder term is obtained under a suitable restriction on the  $\lambda_i$ 's.

### 3. DISTRIBUTION OF A POSITIVE-DEFINITE QUADRATIC FORM

The characteristic function of  $\sum_1^n \lambda_i X_i^2$  is given by

$$\phi(t) = \prod_{j=1}^n (1 - 2i\lambda_j t)^{-\frac{1}{2}}. \quad \dots \quad (1)$$

If we take an arbitrary number  $\lambda$  satisfying

$$\lambda > \frac{1}{2} \max_i \lambda_i \quad \dots \quad (2)$$

the characteristic function may be written

$$\phi(t) = (1 - 2i\lambda t)^{-\frac{n}{2}} \prod_{j=1}^n \left( 1 - \frac{2i\alpha_j t}{1 - 2i\lambda t} \right)^{-\frac{1}{2}} \quad \dots \quad (3)$$

where

$$\alpha_j = \lambda_j - \lambda \quad \dots \quad (4)$$

and it may be expanded as

$$\phi(t) = \sum_{k=0}^{\infty} a_k (-2it)^k (1 - 2i\lambda t)^{-\frac{n}{2} - k}. \quad \dots \quad (5)$$

Here  $a_k$  is the coefficient of  $r^k$  in

$$\prod_{j=1}^n \sum_{i=1}^{\infty} \alpha_j^i \beta_i r^i$$

and

$$\beta_i = \left( -\frac{1}{4} \right)^i \binom{2i}{i}.$$

On applying the inversion formula (cf. Gurland, 1948)

$$F(x) = \frac{1}{2} - \frac{1}{2\pi i} \oint \frac{\phi(t)e^{-itx}}{t} dt \quad \dots \quad (6)$$

and integrating term-by-term it follows\* that the distribution function  $F(x)$  of  $\sum_{i=1}^n \lambda_i X_i^2$  is given by

$$F(x) = \frac{1}{2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)} \int_0^{\frac{x}{\lambda}} v^{\frac{n}{2}-1} e^{-\frac{v}{\lambda}} dv + \sum_{k=1}^{\infty} a_k \frac{\Gamma(k)}{\Gamma\left(k + \frac{n}{2}\right)} \frac{e^{-\frac{x}{\lambda}} x^{\frac{n}{2}} L_{k-1}^{(\frac{n}{2})}\left(\frac{x}{\lambda}\right)}{2^{\frac{n}{2}} \lambda^{k+\frac{n}{2}}} \dots \quad (7)$$

where  $L_m^{(\gamma)}(x)$  is the Laguerre polynomial defined by

$$\left(\frac{d}{dx}\right)^m e^{-x} x^{m+\gamma} = m! e^{-x} x^{\gamma} L_m^{(\gamma)}(x). \quad \dots \quad (8)$$

An equivalent expansion which involves only the  $\chi^2$  distribution explicitly is obtainable as follows:

$$\begin{aligned} & - \frac{1}{2\pi i} \oint \left(\frac{-2it}{1-2i\lambda t}\right)^k \frac{(1-2i\lambda t)^{-\frac{n}{2}}}{t} e^{-itx} dt \\ & = \frac{-1}{2\pi i \lambda^k} \oint \sum_{j=0}^k (-1)^j \binom{k}{j} (1-2i\lambda t)^{-\frac{n}{2}-j} \frac{e^{-itx}}{t} dt. \quad \dots \quad (9) \end{aligned}$$

But 
$$\frac{1}{2} - \frac{1}{2\pi i} \int (1-2it)^{-\frac{f}{2}} \frac{e^{-itx}}{t} dt = G_f(x)$$

the distribution function of  $\chi^2$  with  $f$  degrees of freedom, that is,

$$G_f(x) = \frac{1}{2^{\frac{f}{2}} \Gamma\left(\frac{f}{2}\right)} \int_0^{\frac{x}{2}} e^{-\frac{u}{2}} u^{\frac{f}{2}-1} du.$$

Consequently

$$F(x) = \frac{1}{2} + \sum_{k=0}^{\infty} \frac{a_k}{\lambda^k} \sum_{j=0}^k (-1)^j \binom{k}{j} G_{n+2j}\left(\frac{x}{\lambda}\right). \quad \dots \quad (10)$$

In the notation for the binomial coefficient it is understood here that

$$\binom{k}{0} = 1, \quad k = 0, 1, 2, \dots$$

A bound for the error in approximating to the distribution function  $F(x)$  by partial sums of (7) or (10) will now be developed. The partial sums considered will correspond to  $k = p$  in these series. Thus the remainder term  $R_p(x)$  is given by

$$R_p(x) = F(x) - \frac{1}{2} + \frac{1}{2\pi i} \sum_{k=0}^p a_k \oint \frac{(-2it)^k (1-2i\lambda t)^{-\frac{n}{2}-k}}{t} e^{-itx} dt. \quad \dots \quad (11)$$

\* A detailed proof of (7) is given in Gurland (1955).

Let 
$$r = \max_j \left| \frac{-2it\alpha_j}{1-2i\lambda t} \right| \dots (12)$$

Then from (3) it is clear that

$$|\phi(t)| \leq |1-2i\lambda t|^{-\frac{n}{2}} (1-r)^{-\frac{n}{2}}. \dots (13)$$

Now the Maclaurin series for  $(1-r)^{-\frac{n}{2}}$  yields

$$(1-r)^{-\frac{n}{2}} = 1 + \sum_{j=1}^p \frac{r^j}{jB\left(\frac{n}{2}, j\right)} + \frac{1}{(p+1)B\left(\frac{n}{2}, p+1\right)} \frac{r^{p+1}}{(1-\theta r)^{\frac{n}{2}+p+1}} \dots (14)$$

where  $\theta$  is some number satisfying  $0 < \theta < 1$ . Hence

$$|R_p(x)| \leq \frac{1}{\pi(p+1)B\left(\frac{n}{2}, p+1\right)} \int_0^\infty \frac{|1-2i\lambda t|^{-\frac{n}{2}}}{t} \frac{r^{p+1}}{(1-r)^{\frac{n}{2}+p+1}} dt \dots (15)$$

Now 
$$r = \frac{2\alpha t}{\sqrt{1+4\lambda^2 t^2}}, \quad t > 0 \dots (16)$$

where 
$$\alpha = \max_j |\alpha_j|. \dots (17)$$

On applying the transformation (16) to the integral in (15) we obtain

$$|R_p(x)| \leq \frac{1}{\pi(p+1)B\left(\frac{n}{2}, p+1\right)} \int_0^\delta \frac{u^p \left(1 - \frac{u^2}{\delta^2}\right)^{\frac{n}{2}-1}}{(1-\theta u)^{\frac{n}{2}+p+1}} du \dots (18)$$

where 
$$\delta = \frac{\alpha}{\lambda}. \dots (19)$$

Finally, an application of the mean-value theorem of the integral calculus enables us to write

$$|R_p(x)| < \frac{\delta^{p+1} B\left(\frac{p+1}{2}, \frac{n}{4}\right)}{\pi(p+1)(1-\delta)^{\frac{n}{2}+p+1} B\left(p+1, \frac{n}{2}\right)} \dots (20)$$

By means of the duplication formula for Gamma functions

$$\Gamma(2z) = \frac{2^{2z-1} \Gamma(z) \Gamma\left(z + \frac{1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)} \dots (21)$$

it can be shown that

$$\frac{B\left(\frac{p+1}{2}, \frac{n}{4}\right)}{B\left(p+1, \frac{n}{2}\right)} = \frac{2\sqrt{\pi} \Gamma\left(\frac{n+2p+2}{4}\right)}{\Gamma\left(\frac{p+2}{2}\right) \Gamma\left(\frac{n+2}{4}\right)} \dots (22)$$



Consequently

$$|R_p(x)| < \frac{2}{(p+1)\sqrt{\pi}} \frac{\delta^{p+1}}{(1-\delta)^{\frac{n}{2}+p+1}} \frac{\Gamma\left(\frac{n+2p+2}{4}\right)}{\Gamma\left(\frac{p+2}{2}\right)\Gamma\left(\frac{n+2}{4}\right)} = A_p, \text{ say. } \dots (23)$$

Since

$$\frac{\Gamma\left(\frac{2p+2}{4} + n\right)}{\Gamma\left(\frac{2p+2}{4}\right)} \sim \left(\frac{2p+2}{4}\right)^n, \quad p \rightarrow \infty \dots (24)$$

as is evident from Stirling's formula, it follows that

$$\lim_{p \rightarrow \infty} A_p = 0 \dots (25)$$

provided

$$\frac{\delta}{1-\delta} < 1. \dots (26)$$

This implies that  $\lambda$  must satisfy

$$\frac{2}{3} \max_j \lambda_j < \lambda < 2 \min_j \lambda_j. \dots (27)$$

Hence a suitable value of  $\lambda$  can be found if the  $\lambda_j$ 's are close enough to satisfy

$$\min_j \lambda_j > \frac{1}{3} \max_j \lambda_j. \dots (28)$$

It should be noted, however, that

$$\lim_{p \rightarrow \infty} R_p(x) = 0 \dots (29)$$

for every  $\lambda$  satisfying (2), and that the additional restriction (27) is required to ensure that the uniform bound  $A_p$  approaches zero.

## 4. DISTRIBUTION OF AN INDEFINITE QUADRATIC FORM

Suppose the quadratic form is

$$\sum_1^{n_1} \lambda_i X_i^2 - \sum_{n_1+1}^n \lambda_i X_i^2 \quad \dots \quad (30)$$

where

$$n = n_1 + n_2$$

and

$$\lambda_i > 0, \quad i = 1, 2, \dots, n.$$

Define  $\alpha_j$  and  $\lambda$  as in § 3 and write the characteristic function as

$$\phi(t) = (1-2it\lambda)^{-\frac{n_1}{2}} (1+2it\lambda)^{-\frac{n_2}{2}} \prod_{j=1}^{n_1} \left(1 - \frac{2it\alpha_j}{1-2it\lambda}\right)^{-\frac{1}{2}} \prod_{j=n_1+1}^n \left(1 + \frac{2it\alpha_j}{1+2it\lambda}\right)^{-\frac{1}{2}} \quad \dots \quad (31)$$

This may be expanded as a product of  $n$  power series to yield

$$\phi(t) = \sum_{k=0}^{\infty} \sum_{j=0}^{\infty} (-1)^j a_j b_k (2it)^k (1-2it\lambda)^{-\frac{n_1}{2}-j} (1+2it\lambda)^{-\frac{n_2}{2}-k} \quad \dots \quad (32)$$

where  $a_j$  and  $b_k$  are expressible as in § 3.

The inversion formula may be applied to this expanded form of the characteristic function and a convergent series\* obtained for the distribution function  $F(x)$  of (30). An equivalent but possibly simpler expansion can be obtained by noting that

$$\begin{aligned} & -\frac{1}{2\pi i} \oint \left(\frac{-2it}{1-2i\lambda t}\right)^j \left(\frac{2it}{1+2i\lambda t}\right)^k (1-2i\lambda t)^{-\frac{n_1}{2}} (1+2i\lambda t)^{-\frac{n_2}{2}} e^{-itx} dt \\ &= -\frac{1}{2\pi i \lambda^{j+k}} \sum_{r=0}^j \sum_{s=0}^k (-1)^{r+s} \binom{j}{r} \binom{k}{s} \oint \frac{(1-2i\lambda t)^{-\frac{n_1}{2}-r} (1+2i\lambda t)^{-\frac{n_2}{2}-s}}{t} e^{-itx} dt \quad \dots \quad (33) \end{aligned}$$

and

$$\frac{1}{2} - \frac{1}{2\pi i} \oint \frac{(1-2i\lambda t)^{-\frac{n_1}{2}-r} (1+2i\lambda t)^{-\frac{n_2}{2}-s}}{t} e^{-itx} dt = H_{n_1+2r, n_2+2s} \left(\frac{x}{\lambda}\right) \quad \dots \quad (34)$$

say, where  $H_{f_1, f_2}(x)$  is the distribution function of the difference of two independent  $\chi^2$  random variables with degrees of freedom  $f_1, f_2$  respectively.

\*This series is developed in Gurland (1955).

# QUADRATIC FORMS IN NORMALLY DISTRIBUTED RANDOM VARIABLES

In Gurland (1955) it is shown, under the assumption that  $f_1$  is even,

$$H_{f_1, f_2}(x) = \begin{cases} K + \frac{1}{c} \sum_{h=0}^{\frac{f_1}{2}-1} \binom{\frac{f_1}{2}-1}{h} \Gamma\left(h + \frac{f_2}{2}\right) \int_0^x e^{-\frac{v}{2}} \frac{f_1}{v^{\frac{f_1}{2}-1-h}} dv, & x > 0 \\ \frac{1}{c} \sum_{h=0}^{\frac{f_1}{2}-1} \binom{\frac{f_1}{2}-1}{h} \int_{-\infty}^x e^{-\frac{v}{2}} \frac{f_1}{v^{\frac{f_1}{2}-1-h}} dv \int_{-v}^{\infty} e^{-y} y^{h+\frac{f_2}{2}-1} dy, & x < 0 \end{cases} \quad \dots (35)$$

where

$$K = \frac{1}{B\left(\frac{f_1}{2}, \frac{f_2}{2}\right)^{\frac{1}{2}}} \int_0^1 (1-w)^{\frac{f_1}{2}-1} w^{\frac{f_2}{2}-1} dw \quad \dots (36)$$

and

$$c = 2^{\frac{f_1+f_2}{2}} \Gamma\left(\frac{f_1}{2}\right) \Gamma\left(\frac{f_2}{2}\right). \quad \dots (37)$$

The repeated integral which appears above, for the case  $x < 0$ , can be simplified, under the additional assumption that  $f_2$  is even, as follows

$$\begin{aligned} \int_{-\infty}^x e^{-\frac{v}{2}} \frac{f_1}{v^{\frac{f_1}{2}-1-h}} dv \int_{-v}^{\infty} e^{-y} y^{h+\frac{f_2}{2}-1} dy &= \int_{-\infty}^x e^{-\frac{v}{2}} \frac{f_1}{v^{\frac{f_1}{2}-1-h}} dv \int_0^{\infty} e^{-(z-v)} (z-v)^{h+\frac{f_2}{2}-1} dz \\ &= \sum_{j=0}^{h+\frac{f_2}{2}-1} (-1)^j \binom{h+\frac{f_2}{2}-1}{j} \Gamma\left(h+\frac{f_2}{2}-j\right) \int_{-\infty}^x e^{\frac{v}{2}} v^{j+\frac{1}{2}-1-h} dv, \quad \dots (38) \end{aligned}$$

and this can be evaluated from tables of the Incomplete Gamma Function.

If the inversion formula is applied to (32) and the result in (33) is utilized, the following expansion for the distribution function  $F(x)$  of the indefinite quadratic form obtains:

$$F(x) = \lim_{p \rightarrow \infty} S_p(x) \quad \dots (39)$$

where

$$S_p(x) = \sum_{k=0}^p \sum_{j=0}^p \frac{a_j b_k}{\lambda^{j+k}} \sum_{r=0}^j \sum_{s=0}^k (-1)^{r+s} \binom{j}{r} \binom{k}{s} H_{n_1+2r, n_2+2s}\left(\frac{x}{\lambda}\right). \quad \dots (40)$$



The remainder term of the series will now be considered. Let

$$R_p(x) = F(x) - S_p(x). \quad \dots (41)$$

As in § 3 define

$$r_j = \alpha_j \frac{-2it}{1-2i\lambda t}$$

and

$$r = \max_{1 \leq j \leq n} |r_j|.$$

Then

$$|\phi(t)| \leq |1-2i\lambda t|^{-\frac{n_1}{2}} |1+2i\lambda t|^{-\frac{n_2}{2}} (1-r)^{-\frac{n}{2}} \quad \dots (42)$$

and

$$\begin{aligned} |R_p(x)| \leq & \frac{1}{\pi(p+1)} \int_0^\infty \frac{|1-2i\lambda t|^{-\frac{n}{2}}}{(1-r)^{\frac{n}{2}+p+1}} \left[ r^{p+1} \left\{ \frac{1}{B\left(\frac{n_1}{2}, p+1\right)} + \frac{1}{B\left(\frac{n_2}{2}, p+1\right)} \right\} + \right. \\ & \left. + \frac{r^{2p+2}}{\pi(p+1)B\left(\frac{n_1}{2}, p+1\right)B\left(\frac{n_2}{2}, p+1\right)} \right] dt \quad \dots (43) \end{aligned}$$

as is evident from (15).

On applying the result of (23) we may write

$$\begin{aligned} |R_p(x)| < \frac{1}{\pi(p+1)} \left[ \frac{2\delta^{p+1}\sqrt{\pi}}{\Gamma(p+2)} \left\{ \frac{\Gamma\left(\frac{n_1+2p+2}{4}\right)}{(1-\delta)^{\frac{n_1}{2}+p+1}\Gamma\left(\frac{n_1+2}{4}\right)} + \frac{\Gamma\left(\frac{n_2+2p+2}{4}\right)}{(1-\delta)^{\frac{n_2}{2}+p+1}\Gamma\left(\frac{n_2+2}{4}\right)} \right\} + \right. \\ & \left. + \frac{\delta^{2p+2}B\left(\frac{n}{4}, p+1\right)}{\pi(p+1)(1-\delta)^{\frac{n}{2}+2p+2}B\left(\frac{n_1}{2}, p+1\right)B\left(\frac{n_2}{2}, p+1\right)} \right] \quad \dots (44) \end{aligned}$$

Similar remarks, of course, are relevant here concerning the closeness of the  $\lambda_j$ 's.

# QUADRATIC FORMS IN NORMALLY DISTRIBUTED RANDOM VARIABLES

## 5. DISTRIBUTION OF A RATIO OF QUADRATIC FORMS

Consider the ratio  $\frac{XQX'}{XPX'}$ , where  $Q$  and  $P$  are arbitrary symmetric  $n \times n$  matrices but  $P$  is positive definite. There is no loss of generality in assuming that the random variables  $X_1, X_2, \dots, X_n$  are normal and independent, each with mean zero and variance one. The joint characteristic function of the numerator and denominator is given by

$$\phi(t_1, t_2) = E e^{i(t_1 XQX' + t_2 XPX')} = |I - 2it_1 Q - 2it_2 P|^{-\frac{1}{2}} \quad \dots (45)$$

where  $I$  denotes the unit  $n \times n$  matrix. From the theory of inversion formulae (Gurland 1948) the distribution function  $F(x)$  of the ratio becomes

$$F(x) = \frac{1}{2} - \frac{1}{2\pi i} \oint \sqrt{\frac{1}{|I - 2it(Q - xP)|}} \frac{dt}{t} \quad \dots (46)$$

Since  $Q - xP$  is symmetric its characteristic roots are real; consequently we may write

$$|I - 2it(Q - xP)| = \prod_{j=1}^{n_1} (1 - 2i\lambda_j t) \prod_{j=n_1+1}^n (1 + 2i\lambda_j t) \quad \dots (47)$$

where  $\lambda_j > 0$ ;  $j = 1, 2, \dots, n_1, n_1+1, \dots, n$   
and  $n = n_1 + n_2$ .

From (46) and (47) it is evident that the reasoning which leads to a convergent series for  $F(x)$  closely resembles that of § 4. As before, take

$$\lambda > \frac{1}{2} \max_{1 \leq j \leq n} \lambda_j$$

$$\alpha_j = \lambda_j - \lambda;$$

and let

then

$$\prod_{j=1}^{n_1} (1 - 2i\lambda_j t)^{-\frac{1}{2}} = \sum_{k=0}^{\infty} a_k (-2it)^k (1 - 2i\lambda t)^{-\frac{n_1}{2} - k}, \quad \dots (48)$$

$$\prod_{j=n_1+1}^n (1 + 2i\lambda_j t)^{-\frac{1}{2}} = \sum_{k=0}^{\infty} b_k (2it)^k (1 + 2i\lambda t)^{-\frac{n_2}{2} - k}$$

and the distribution function  $F(x)$  of the ratio may be written as

$$F(x) = \frac{1}{2} - \frac{1}{2\pi i} \lim_{p \rightarrow \infty} \sum_{k=0}^p \sum_{j=0}^p (-1)^j a_j b_k \oint \frac{(2it)^{j+k}}{(1-2i\lambda t)^{\frac{n_1}{2}+j} (1+2i\lambda t)^{\frac{n_2}{2}+k}} \frac{dt}{t} \dots (49)$$

The integral occurring in the generic term may be simplified by noting that

$$\begin{aligned} & -\frac{1}{2\pi i} \oint \left( \frac{-2it}{1-2i\lambda t} \right)^j \left( \frac{2it}{1+2i\lambda t} \right)^k \frac{(1-2i\lambda t)^{-\frac{n_1}{2}} (1+2i\lambda t)^{-\frac{n_2}{2}}}{t} dt \\ &= -\frac{1}{2\pi i \lambda^{j+k}} \sum_{r=0}^j \sum_{s=0}^k (-1)^{r+s} \binom{j}{r} \binom{k}{s} \oint \frac{(1-2i\lambda t)^{-\frac{n_1}{2}-r} (1+2i\lambda t)^{-\frac{n_2}{2}-s}}{t} dt \dots (50) \end{aligned}$$

and

$$\begin{aligned} & \frac{1}{2} - \frac{1}{2\pi i} \oint \frac{(1-2i\lambda t)^{-\frac{f_1}{2}} (1+2i\lambda t)^{-\frac{f_2}{2}}}{t} dt \\ &= \frac{1}{B\left(\frac{f_1}{2}, \frac{f_2}{2}\right)} \int_0^1 \frac{u^{\frac{f_1}{2}-1}}{(1+u)^{\frac{f_1+f_2}{2}}} du \\ &= \frac{1}{B\left(\frac{f_1}{2}, \frac{f_2}{2}\right)} \int_{\frac{1}{2}}^1 (1-v)^{\frac{f_1}{2}-1} v^{\frac{f_2}{2}-1} dv = 1 - I_{f_1, f_2}\left(\frac{1}{2}\right), \dots (51) \end{aligned}$$

say, where  $I_{f_1, f_2}(x)$  is the Incomplete Beta function evaluated at  $x$ . Consequently the following expansion for the distribution function  $F(x)$  of the ratio obtains

$$F(x) = \frac{1}{2} + \lim_{p \rightarrow \infty} \sum_{k=0}^p \sum_{j=0}^p \frac{a_j b_k}{\lambda^{j+k}} \sum_{r=0}^j \sum_{s=0}^k (-1)^{r+s+1} \binom{j}{r} \binom{k}{s} I_{n_1+2r, n_2+2s}\left(\frac{1}{2}\right) \dots (52)$$

By applying the same reasoning as in § 4, the remainder  $R_p(x)$  is seen to satisfy the same inequality (44). This bound, however, is not uniform for the present case since  $\delta_p$  depends on the value  $x$  here.



# QUADRATIC FORMS IN NORMALLY DISTRIBUTED RANDOM VARIABLES

## 6. EXPECTATION OF A RATIO OF QUADRATIC FORMS

Let 
$$M = E \frac{XQX'}{XPX'}$$

and assume the same notation as in § 5. Denote the joint moment-generating function of the numerator and denominator by

$$\psi(t_1, t_2) = Ee^{(t_1XQX' + t_2XPX')}$$

where  $t_1, t_2$  are real. Then

$$\psi(t_1, t_2) = |I - 2t_1Q - 2t_2P|^{-1}.$$

Now the  $k$ -th moment of  $\frac{XQX'}{XPX'}$  is given by

$$\int_{-\infty}^0 \int_{-\infty}^0 \dots \int_{-\infty}^0 \left\{ \frac{\partial^k \psi}{\partial t_1^k} \right\}_{t_1=0} dt_{21} dt_{22} \dots dt_{2k} \quad \dots \quad (53)$$

where

$$t_2 = t_{21} + t_{22} + \dots + t_{2k}.$$

The present article considers only the case  $k = 1$  although the method described below applies generally. Let

$$\frac{\partial}{\partial t_1} \left\{ |I - 2t_1Q - 2t_2P| \right\}_{t_1=0} = \sum_{j=0}^{n-1} c_j t_2^j \quad \dots \quad (54)$$

and

$$|I - 2t_2P| = \prod_{j=1}^n (1 - 2t_2 v_j). \quad (55)$$

On applying (53) with  $k = 1$  it follows that

$$M = -\frac{1}{2} \int_{-\infty}^0 \left\{ \sum_{j=0}^{n-1} c_j t^j \right\} \left\{ \prod_{j=1}^n (1 - 2tv_j)^{-\frac{3}{2}} \right\} dt. \quad \dots \quad (56)$$

Now

$$(1 - 2tv_j)^{-\frac{3}{2}} = (1 - 2tv)^{-\frac{3}{2}} \sum_{k=0}^{\infty} \gamma_k \left( \frac{-2t\alpha_j}{1 - 2tv} \right)^k \quad \dots \quad (57)$$

where

$$v > \frac{1}{2} \max_j v_j,$$

$$\alpha_j = v_j - v,$$

$$\gamma_k = (-1)^k \frac{\Gamma(\frac{3}{2} + k)}{\Gamma(\frac{3}{2})}.$$

Consequently

$$\prod_{j=1}^n (1-2tv_j)^{-\frac{3}{2}} = (1-2tv)^{-\frac{3n}{2}} \sum_{k=0}^{\infty} g_k (-2t)^k (1-2tv)^{-k} \quad \dots (58)$$

where  $g_k$  is the coefficient of  $r^k$  in

$$\prod_{j=1}^n \sum_{k=1}^{\infty} \gamma_k (\alpha_j r)^k.$$

It should be remarked here that since the  $g_k$ 's are symmetric functions of the roots  $v_1, v_2, \dots, v_n$ , it is not necessary to know these roots explicitly.

In virtue of (58) which is uniformly convergent, the following expansion for  $M$  can be derived from (56)

$$M = -\frac{1}{2} \sum_{j=0}^{n-1} \sum_{k=0}^{\infty} c_j g_k \int_{-\infty}^0 \frac{t^j (-2t)^k}{(1-2tv)^{\frac{3n}{2}+k}} dt. \quad \dots (59)$$

But

$$\begin{aligned} \int_{-\infty}^0 \frac{t^{j+k}}{(1-2tv)^{\frac{3n}{2}+k}} dt &= \frac{(-1)^{j+k}}{(2v)^{j+k+1}} \int_0^{\infty} \frac{t^{j+k} dt}{(1+t)^{\frac{3n}{2}+k}} \\ &= \frac{(-1)^{j+k}}{(2v)^{j+k+1}} B(j+k+1, \frac{3n}{2} - j - 1). \end{aligned} \quad \dots (60)$$

Thus

$$M = \sum_{j=0}^{n-1} \sum_{k=0}^{\infty} \frac{(-1)^{j+1}}{2^{j+2} v^{j+k+1}} c_j g_k B(j+k+1, \frac{3n}{2} - j - 1). \quad \dots (61)$$

We now proceed to find a bound on the remainder term which arises in taking a partial sum of the above expansion as an approximation to  $M$ . Denote the generic term of (61) by

$$u_{jk} = \frac{(-1)^{j+1}}{2^{j+2} v^{j+k+1}} c_j g_k B(j+k+1, \frac{3n}{2} - j - 1) \quad \dots (62)$$

and write

$$R_p = \sum_{j=0}^{n-1} \sum_{k=0}^{\infty} u_{jk} - \sum_{j=0}^{n-1} \sum_{k=0}^p u_{jk} \quad \dots (63)$$

and

$$R_{j,p} = \sum_{k=0}^{\infty} u_{jk} - \sum_{k=0}^p u_{jk}. \quad \dots (64)$$

# QUADRATIC FORMS IN NORMALLY DISTRIBUTED RANDOM VARIABLES

First we consider  $R_{jp}$  and note that

$$|R_{jp}| \leq \int_{-\infty}^0 |c_j| |t|^j |1-2vt|^{-\frac{3n}{2}} \sum_{k=p+1}^{\infty} |g_k| \left| \frac{-2t}{1-2vt} \right|^k dt \quad \dots (65)$$

Let 
$$r = \max_j \left| \frac{2\alpha_j t}{1-2vt} \right| = \frac{2\alpha |t|}{|1-2vt|} \quad \dots (66)$$

where 
$$\alpha = \max_j |\alpha_j|.$$

Then 
$$\prod_{j=1}^n |1-2vt_j|^{-\frac{3}{2}} \leq |1-2vt|^{-\frac{3n}{2}} (1-r)^{-\frac{3n}{2}} \quad \dots (67)$$

and 
$$\left| \sum_{k=p+1}^{\infty} g_k \left( \frac{-2t}{1-2vt} \right)^k \right| \leq \frac{1}{(p+1)B\left(\frac{3n}{2}, p+1\right)} \frac{r^{p+1}}{(1-\theta r)^{\frac{3n}{2}+p+1}} \quad \dots (68)$$

by reasoning as in § 3. Hence

$$|R_{jp}| \leq \frac{1}{(p+1)B\left(\frac{3n}{2}, p+1\right)} \int_{-\infty}^0 \frac{r^{p+1}}{|1-2vt|^{\frac{3n}{2}}} \frac{|t|^j |c_j|}{(1-r)^{\frac{3n}{2}+p+1}} dt. \quad \dots (69)$$

The transformation

$$w = \frac{-2\alpha t}{1-2vt} \quad \dots (70)$$

reduces the integral in (69) to

$$\frac{|c_j|}{(2\alpha)^{j+1}} \int_0^{\delta} \frac{w^{p+j+1}}{(1-w)^{\frac{3n}{2}+p+1}} \left(1 - \frac{w}{\delta}\right)^{\frac{3n}{2}-j-2} dw. \quad \dots (71)$$

From this it readily follows that

$$|R_{jp}| < \frac{|c_j| \delta^{p+j+2} B\left(p+j+2, \frac{3n}{2} - j - 1\right)}{(p+1)B\left(\frac{3n}{2}, p+1\right) (2\alpha)^{j+1} (1-\delta)^{\frac{3n}{2}+p+1}}. \quad \dots (72)$$



This yields a bound for  $R_p$  since

$$|R_p| < \sum_{j=0}^{n-1} |R_{jp}|. \quad \dots (73)$$

## REFERENCES

- BHATTACHARYA, A. (1945): A note on the distribution of the sum of chi-squares. *Sankhyā*, **7**, 27-28.
- GURLAND, J. (1948): Inversion formulae for the distribution of ratios. *Ann. Math. Stat.*, **19**, 228-237.
- (1953): Distribution of quadratic forms and ratios of quadratic forms. *Ann. Math. Stat.*, **24**, 416-427.
- (1956): Distribution of definite and of indefinite quadratic forms. *Ann. Math. Stat.*, **26**, 122-127.
- ROBBINS, H. and PITMAN, E. J. G. (1949): Application of the method of mixtures to quadratic forms in normal variates. *Ann. Math. Stat.*, **20**, 552-560.

*Paper Received: October, 1954.*

# A METHOD OF DISCRIMINATION IN TIME SERIES ANALYSIS—II

By A. RUDRA

*Indian Statistical Institute, Calcutta*

## 1. INTRODUCTION

1.1 This part of the paper is intended to give fuller discussion of certain points regarding the  $F$ -diagram method which was developed in part I of the paper. First we take up for closer study the subject of the configurations of the  $\pi$ -diagram† for Linear Cyclical series; next we consider the question of the distribution of the  $F$ -statistic; finally, we briefly discuss the computational aspect of the method and suggest a method of dealing with fractional periodicities.

## 2. CONFIGURATIONS OF THE $\pi$ -DIAGRAM

2.1. It was mentioned in § 3 of part I that the actual appearance of the  $F$ -diagram is influenced by the length of the series as well as the relation the true periodicities with other trial periods. We shall in the following probe this point, first with reference to series with a single periodicity and then with reference to series will all be different, having several superposed periodicities.

2.2. Let  $p_0$  be as before the true periodicity and  $p$  any trial value relatively prime to  $p_0$ . Then the first  $p_0$  elements that will enter any column in the  $M$  table for  $p$  will all be different, being, but for  $\Delta$ , the different elements that constitute the cycle:

$$C(p_0) = \{\theta_1(p_0), \theta_2(p_0), \dots, \theta_{p_0}(p_0)\}. \quad \dots (2.1)$$

As a result, the elements in each column will constitute a periodic series with the period  $p_0$  and the same elements as constitute  $C(p_0)$  but coming in different orders in the different columns. The following two examples will make the situation clear:

*Example 1:*

$$p = 5; \quad p_0 = 4; \quad C(p_0) = \{\theta_1, \theta_2, \theta_3, \theta_4\}.$$

columns	(1)	(2)	(3)	(4)	(5)
	$\theta_1$	$\theta_2$	$\theta_3$	$\theta_4$	$\theta_1$
	$\theta_2$	$\theta_3$	$\theta_4$	$\theta_1$	$\theta_2$
	$\theta_3$	$\theta_4$	$\theta_1$	$\theta_2$	$\theta_3$

and so on.

---

† All through this paper we shall make use of terminology and symbolism introduced in Part I (*Sankhyā*, 15, 9-34) without repeating the explanations.

Example 2:

$$p = 3; \quad p_0 = 4; \quad C(p_0) = \{\theta_1, \theta_2, \theta_3, \theta_4\}$$

columns	(1)	(2)	(3)
	$\theta_1$	$\theta_2$	$\theta_3$
	$\theta_4$	$\theta_1$	$\theta_2$
	$\theta_3$	$\theta_4$	$\theta_1$
	$\theta_2$	$\theta_3$	$\theta_4$

and so on.

2.3. Let  $p$  be not relatively prime to  $p_0$  and let  $L(pp_0)$  be their least common multiple. The elements in each column in the  $M$ -table for  $p$  will again, but for  $\Delta$ , constitute a periodic series, but now the periodicity will be  $q(p) = \frac{L(pp_0)}{p}$ , different combinations of  $q(p)$  elements from the  $p_0$  elements of  $C(p_0)$  going to the different columns. The following two examples illustrate the situation:

Example 3:

$$p = 6; p_0 = 4; \quad L(pp_0) = 12; \quad q(p) = 2; \quad C(p_0) = \{\theta_1, \theta_2, \theta_3, \theta_4\}.$$

columns	(1)	(2)	(3)	(4)	(5)	(6)
	$\theta_1$	$\theta_2$	$\theta_3$	$\theta_4$	$\theta_1$	$\theta_2$
	$\theta_3$	$\theta_4$	$\theta_1$	$\theta_2$	$\theta_3$	$\theta_4$
	$\theta_1$	$\theta_2$	$\theta_3$	$\theta_4$	$\theta_1$	$\theta_2$

and so on.

Example 4:

$$p = 4; \quad p_0 = 6; L(pp_0) = 12; \quad q(p) = 3; \quad C(p_0) = \{\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6\}.$$

columns	(1)	(2)	(3)	(4)
	$\theta_1$	$\theta_2$	$\theta_3$	$\theta_4$
	$\theta_5$	$\theta_6$	$\theta_1$	$\theta_2$
	$\theta_3$	$\theta_4$	$\theta_5$	$\theta_6$
	$\theta_1$	$\theta_2$	$\theta_3$	$\theta_4$

and so on.



## A METHOD OF DISCRIMINATION IN TIME SERIES ANALYSIS—II

2.4. Let  $g_i(p)$  be the mean of the distinct elements that enter the  $i$ -th column ( $i = 1, 2, \dots, p_0$ ). Thus, in example 3,  $g_1(p) = \frac{\theta_1 + \theta_3}{2}$ ;  $g_2(p) = \frac{\theta_2 + \theta_4}{2}$ ;  $g_3(p) = \frac{\theta_1 + \theta_3}{2}$ ;  $g_4(p) = \frac{\theta_2 + \theta_4}{2}$  and so on. In example 4,  $g_1(p) = \frac{\theta_1 + \theta_5 + \theta_3}{3}$ ;  $g_2(p) = \frac{\theta_2 + \theta_6 + \theta_4}{3}$ ; and so on. Let  $r$  be the number of complete cycles  $C(p_0)$  in the given series of length  $N$  and  $h(p)$  the number of complete cycles in each column of the  $M$ -table for  $p$ . (We assume that there are equal number of elements in each column); that is,

$$N = rp_0 + t_0 \text{ where } t_0 < p_0, \text{ and } \frac{N}{p} = n(p) = h(p) \text{ of } (p) + u(p) \text{ where } u(p) < q(p).$$

We shall use the symbols  $\beta^N, \lambda^N(p), \pi^N(p)$  and  $\eta^N(p)$  to denote the functions  $\beta, \lambda(p), \pi(p)$  and  $\eta(p)$  for a given series of length  $N$ .

Thus

$$\begin{aligned} \beta^N = \frac{1}{N-1} \sum_{ij} \frac{\{m_{ij}(p) - m_{..}\}^2}{\sigma^2} &= \frac{1}{N-1} \left\{ r(p_0-1)\Omega(p_0) + \frac{[\theta_1^2(p_0) + \theta_2^2(p_0) + \dots + \theta_{t_0}^2(p_0)]}{\sigma^2} \right. \\ &\quad \left. - \frac{N}{\sigma^2} \left[ \frac{\theta_1(p_0) + \theta_2(p_0) + \dots + \theta_{t_0}(p_0)}{N} \right]^2 \right\} \quad \dots (2.2) \end{aligned}$$

so that, as  $N \rightarrow \infty$ ,  $\beta^N \rightarrow \frac{p_0-1}{(p_0)} \Omega(p_0)$ .

$$\begin{aligned} \lambda^N(p) = \frac{1}{p-1} \sum_{i=1}^p \frac{\{m_{i.}(p) - m_{..}\}^2}{\sigma^2} \quad n(p) &= \frac{1}{\sigma^2(p-1)} \left\{ \sum_{i=1}^p \frac{[n(p) - u(p)]^2 g_i^2(p)}{n(p)} + \right. \\ &\quad \left. + \sum_{i=1}^p \frac{\phi_i^2(p)}{n(p)} + \sum_{i=1}^p \frac{2\phi_i(p)g_i(p)[n(p) - u(p)]}{n(p)} - N \left[ \frac{\theta_1(p_0) + \dots + \theta_{t_0}(p_0)}{N} \right]^2 \right\} \\ &\dots (2.3) \end{aligned}$$

where  $\phi_i(p) = m_{i.}(p) n(p) - h(p)q(p)g_i(p)$ .

(a) If  $p$  and  $p_0$  are mutually prime,  $g_i(p) = 0$  for all  $i$  so that

$$\lambda^N(p) = \frac{1}{(p-1)\sigma^2} \left\{ \sum_{i=1}^p \frac{\phi_i^2(p)}{n(p)} - N \left[ \frac{\theta_1(p_0) + \dots + \theta_{t_0}(p_0)}{N} \right]^2 \right\} \quad \dots (2.4)$$

so that it tends to zero as  $N$  tends to infinity.

If any given value of  $N$  happens to be a common multiple of  $p$  and  $p_0$ , then in addition to  $g_i(p)$  being zero for all  $i$ ,  $\phi_i(p)$  is also zero for all  $i$  and so is  $t_0$ .  $\lambda^N(p)$  therefore identically vanishes for such a combination of  $p, p_0$  and  $N$ .

It follows that when  $p$  and  $p_0$  are mutually prime,  $\pi^N(p)$  takes the value  $\frac{1}{1 + \frac{N-1}{N-p} \beta^N}$  for all values of  $N$  which are common multiples of  $p$  and  $p_0$ ; and that

it tends to the value of  $\frac{1}{1 + \frac{p_0-1}{p_0} \Omega(p_0)}$  as  $N \rightarrow \infty$ .

(b) If  $p \neq sp_0$  ( $s = 1, 2, \dots$ ) and  $p$  and  $p_0$  are not mutually prime  $\frac{\lambda^N(p)}{N}$  tends to the limit  $\sum_{i=1}^p \frac{g_i^2(p)}{\sigma^2 p(p-1)}$  as  $N$  tends to infinity. Hence, from the relation

$$\beta^N = \frac{p-1}{N-1} \lambda^N(p) + \frac{N-p}{N-1} \eta^N(p) \quad \dots (2.5)$$

we conclude that  $\eta^N(p)$  tends, as  $N$  tends to infinity, to

$$\frac{p_0-1}{p_0} \Omega(p_0) - \frac{1}{p} \sum_{i=1}^p \frac{g_i^2(p)}{\sigma^2} \quad \dots (2.6)$$

and  $\frac{1+\lambda^N(p)}{1+\eta^N(p)}$  is asymptotically equal to  $N \frac{\frac{1}{p(p-1)} \sum_{i=1}^p \frac{g_i^2(p)}{\sigma^2}}{1 + \frac{p_0-1}{p_0} \Omega(p_0) - \frac{1}{p} \sum_{i=1}^p \frac{g_i^2(p)}{\sigma^2}}$  ... (2.7)

(c) If  $p = sp_0$  for some positive integral  $s$ ,

$$\pi^N(p) = 1 + \frac{N-1}{sp_0-1} \beta^N$$

so that  $\frac{\pi^N(p)}{N}$  tends as  $N$  tends to infinity to  $\frac{p_0-1}{p_0} \frac{\Omega(p_0)}{sp_0-1}$ .  $\pi^N(p)$  therefore is asymptotically equal to  $\frac{p_0-1}{p_0} \frac{\Omega(p_0)}{sp_0-1} N$ .

2.5. Thus, by increasing  $N$  we shall increase all the  $F$  values except those at points relatively prime to  $p_0$ . Increase in sample size is therefore not an unmixed blessing. While the sharpness of peaks at genuine periodicities are thereby increased, the magnitude of some of the spurious peaks may also be affected in the same way. An increase in  $N$  is however advantageous from the point of view of estimation. This

## A METHOD OF DISCRIMINATION IN TIME SERIES ANALYSIS—II

will be referred to in a later section. It should however be noted that what is important is the number of rows and not the actual number of observations. As long as the number of rows is less than  $p_0$ , the difference noted above between values of  $p$  which are relatively prime to  $p_0$  and which are not, will not be observed.

2.6. So far we have been considering the effect of the values of  $p$  and  $N$  in relation to a single periodicity  $p_0$ . The matter becomes very complicated when there are more than one periodicity. Suppose that there are  $k$  periodicities  $p_1, p_2, \dots, p_k$ . Then it is desirable that when the series is arranged in a Buys Ballot table for trial period  $p_i$  ( $i = 1, 2, \dots, k$ ) the elements of the other existing periods  $p_j \neq p_i$  cause the least amount of damping effect on the tendency of the period  $p_i$  to produce a peak in the  $F$ -diagram. The effect of these other periods will of course depend on several factors. Thus, it may be surmised that the damping will be more, the more the relative importance of these other periods compared with  $p_i$ . The relative importance may be measured by  $\Omega(p_j)$  ( $j \neq i$ ). Further, the contribution to the mean within column sum of squares  $\eta(p_i)$  due to a particular period  $p_j$  will be more if  $p_j$  is relatively prime to  $p_i$  than if it is not. Now, the smaller the value of  $\eta(p_i)$ , the greater the magnitude of  $\pi(p_i)$ . Hence, we have a better chance of observing a peak of  $p_i$  if the other periods  $p_j (\neq p_i)$  are not relatively prime to  $p_i$ . On the other hand for such values of  $p$  that do not correspond to any periodicity, it is desirable that  $\eta(p)$  is greater than  $\lambda(p)$  and for this it is better to have as many of the values  $p_i$  ( $i = 1, 2, \dots, k$ ) relatively prime to  $p$  as possible. As to the effect of the sample size  $N$ , the same remarks apply here as in the case of a single periodicity.

2.7. It is however more desirable to have the values  $p_i$  ( $i = 1, 2, \dots, k$ ) relatively prime to each other from the point of estimation. Assume that this is true and suppose that  $N = sp_1p_2 \dots p_k$  where  $s$  is a positive integer. Then, when the series is arranged in a Buys Ballot table for trial period  $p_i$  ( $i = 1, 2, \dots, k$ ) there are  $\frac{N}{p_i}$  rows, and the elements of the cycle  $C(p_j)$  complete  $\frac{N}{p_i p_j}$  complete cycles in each column. As each cycle adds up to zero, the column means do not get any contribution whatsoever from elements belonging to any other period than  $p_i$  and hence

$$m_{t.}(p_i) = \Delta + \theta_t(p_i) \quad (t = 1, 2, \dots, p_i) \quad \dots \quad (2.8)$$

and

$$m_{..} = \Delta$$

The least square estimate of  $\theta_t(p_i)$  is therefore given by

$$\bar{x}_{t.}(p_i) - \bar{x} \dots (t = 1, 2, \dots, p_i) \quad \dots \quad (2.9)$$

This simple solution is not available when  $p_1, p_2, \dots, p_k$  are not mutually prime. Even when they are so,  $N$  would rarely be a common multiple of all of them. We shall therefore in general have

$$m_{t.}(p_i) = \Delta + \theta_t(p_i) + \delta_t(p_i) \quad (t = 1, 2, \dots, p_i) \quad \dots \quad (2.10)$$



If  $p_1 p_2 \dots p_k$  are relatively prime, then, given sufficiently large numbers of rows, the contribution of each periodicity except  $p$ , to any column mean would independently amount to very nearly zero, that is

$$\delta_{i\cdot}(p_i) \simeq 0. \quad \dots (2.11)$$

As  $\Delta$  could in any case approximately estimated by  $\bar{x}_{..}$  we have, for large  $N$ , the following for the approximate estimate of  $\theta_i(p_i)$ :

$$\bar{x}_{i\cdot}(p_i) - \bar{x}_{..} \quad (i = 1, 2, \dots, p_i). \quad \dots (2.12)$$

### 3. DISTRIBUTION OF $F(p)$ FOR A LINEAR CYCLICAL SERIES

3.1. We have seen in § 2 of Part I, that  $F(p)$  for a Linear Cyclical Series is distributed as

$$\frac{\chi^2[\lambda(p)]}{\chi^2[\eta(p)]} \times \frac{N-p}{p-1} \quad \dots (3.1)$$

where  $\chi^2[\lambda(p)]$  and  $\chi^2[\eta(p)]$  are noncentral chi-squares with noncentrality parameters  $(p-1)\lambda(p)$  and  $(N-p)\eta(p)$  respectively. We shall now show that these two noncentral chi-squares are independent so that  $F(p)$  may have the density function given by formula (2.10) of Part I.

Let us consider the following two quadratic relations

$$\sum_{ij} \{y_{ij}(p) - \bar{y}_{..}\}^2 = \sum_{ij} \{y_{ij}(p) - \bar{y}_{i\cdot}(p)\}^2 + \sum_{i=1}^p \{\bar{y}_{i\cdot}(p) - \bar{y}_{..}\}^2 n(p) \quad \dots (3.2)$$

$$\sum_{ij} \{\epsilon_{ij}(p) - \epsilon_{..}\}^2 = \sum_{ij} \{\epsilon_{ij}(p) - \epsilon_{i\cdot}(p)\}^2 + \sum_{i=1}^k \{\epsilon_{i\cdot}(p) - \epsilon_{..}\}^2 n(p). \quad \dots (3.3)$$

There is an orthogonal transformation that, applied to  $y_{ij}(p)$  ( $i = 1, 2, \dots, p; j = 1, 2, \dots, n(p)$ ) reduces relation (3.2) to

$$\sum_{t=1}^{N-1} Z_t^2 = \left\{ \sum_{t=1}^{N-p} Z_t^2 \right\} + \left\{ \sum_{t=N-p+1}^{N-1} Z_t^2 \right\} \quad \dots (3.4)$$

and the same transformation applied to  $\epsilon_{ij}(p)$  ( $i = 1, 2, \dots, p; j = 1, 2, \dots, n(p)$ ) reduces (3.3) to

$$\sum_{t=1}^{N-1} \gamma_t^2 = \left\{ \sum_{t=1}^{N-p} \gamma_t^2 \right\} + \left\{ \sum_{t=N-p+1}^{N-1} \gamma_t^2 \right\}. \quad \dots (3.5)$$

But

$$y_{ij}(p) = m_{ij}(p) + \epsilon_{ij}(p)$$

so that  $Z_t = \zeta_t + \gamma_t$  where  $\zeta_t$  is the same linear function of the  $m_{ij}$ 's as  $Z_t$  and  $\gamma_t$  are of  $y_{ij}(p)$ 's and  $\epsilon_{ij}(p)$ 's. But  $\gamma_t$ 's are independent; hence  $Z_t$ 's are also independent.

Hence  $\sum_{t=1}^{N-p} Z_t^2$  is independent of  $\sum_{t=N-p+1}^{N-1} Z_t^2$ . Hence

$$\chi^2[\lambda(p)] = \sum_{i=1}^p \left\{ \frac{\{\bar{y}_{i.}(p) - \bar{y}_{..}\}^2 n(p)}{\sigma^2} \right\}$$

and  $\chi^2[\eta(p)] = \sum_{ij} \frac{\{y_{ij}(p) - \bar{y}_{i.}(p)\}^2}{\sigma^2}$  are independent.

3.2. The first two moments of  $\frac{\chi^2[\eta(p)]}{N-p}$  are  $1 + \eta(p)$  and  $\frac{2\{1 + 2\eta(p)\}}{N-p}$  respectively, so that as  $N \rightarrow \infty$  the first moment remains constant but the second moment converges to zero. This means that  $\frac{\chi^2[\eta(p)]}{N-p}$  converges to  $1 + \eta(p)$  in probability. The distribution function of  $\chi^2[\lambda(p)]$  is however not affected by the passage of  $N$  to infinity. Hence the distribution function of  $\frac{\chi^2[\lambda(p)]}{\chi^2[\eta(p)]} \times \frac{N-p}{p-1}$  converges to that of  $\frac{\chi^2[\lambda(p)]}{(p-1)(1+\eta(p))}$ .

3.3. We shall now endeavour to justify our practice of arguing as to the occurrence of peaks in the  $F$ -diagram on the basis of occurrence of peaks in the  $\pi$ -diagram. We shall use the result of Patnaik (1949) that  $\chi^2[\lambda(p)]$  can be reasonably approximated to by  $\rho \chi_v^2$  where  $\chi_v^2$  is a central chi-square with the fractional degree of freedom  $v$ ,  $v$  and  $\rho$  being given by

$$\rho = \frac{1 + 2\lambda(p)}{1 + \lambda(p)} \quad \dots (3.6)$$

$$v = (p-1) \frac{\{1 + \lambda(p)\}^2}{1 + 2\lambda(p)}. \quad \dots (3.7)$$

Even for small values of  $p$ ,  $v$  given by (3.8), may, depending on the magnitude of  $\lambda(p)$ , will be sufficiently large to allow of the treatment of  $\sqrt{2}\chi_v^2$  as a normal variable with unit standard deviation and mean equal to  $\sqrt{2}v$ .

Using this approximation we can treat

$$\sqrt{F(p)} = \sqrt{\frac{1+\bar{\lambda}(p)}{1+\eta(p)}} \sqrt{\frac{1+2\lambda(p)}{2\{1+\lambda(p)\}\{1+\eta(p)\}(p-1)}} \quad \dots (3.8)$$

as a unit normal variable.

The probability that  $F(p)$  exceeds a certain quantity  $K$ , is therefore roughly given by the probability of the following inequality:

$$d > \frac{-\sqrt{\frac{1+\lambda(p)}{1+\eta(p)}} + \sqrt{K}}{\sqrt{\frac{1+2\lambda(p)}{2\{1+\lambda(p)\}\{1+\eta(p)\}(p-1)}}} \quad \dots (3.9)$$

where  $d$  is a unit normal variable.

Now

$$\frac{1+\lambda(p)}{1+\eta(p)} = \pi(p)$$

and

$$\frac{1+2\lambda(p)}{2\{1+\bar{\lambda}(p)\}\{1+\eta(p)\}(p-1)} \simeq \frac{1}{(1+\eta(p))(p-1)}$$

so that the right hand side of (3.9) is a decreasing function of  $\pi(p)$ . Hence the probability of  $F(p)$  at a certain point exceeding a given level increases with  $\pi(p)$ .

3.4. While it is difficult to say anything about the probability of the  $F(p)$ -diagram passing beyond a certain level at a certain point as we have not considered the joint distribution of the  $F$ -statistics for different values of  $p$ , we shall show here as a justification of the use of probability curves to define the different regions in the  $F$ -diagram that under moderate assumptions the probability that  $F(p)$  for a linear cyclical series at a point  $p$  not corresponding to any true periodicity passes beyond the  $\alpha\%$  curve is at most  $\alpha\%$ . This we prove by making use of the central chi-square approximation for  $\chi^2[\lambda(p)]$  discussed in the last section but without resorting to the further normal approximation. The assumption we make is that for  $p$  not corresponding to any true periodicity,  $\lambda(p)$  and  $\eta(p)$  differ negligibly from  $\beta$ .

In other words, we treat  $F(p)$  as a Pearsonian variable with the first two moments

$$\mu_1[F(p)] \simeq 1$$

$$\mu_2[F(p)] \simeq \frac{2}{p-1} \frac{1+2\beta}{(1+\beta)^2} \quad \dots (3.10)$$



and the  $\beta$ -ratios

$$\beta_1(p) = \frac{8}{p-1} \frac{\{1+3\beta\}^2}{\{1+2\beta\}^3} \quad \dots (3.11)$$

and

$$\beta_2(p) = 3 + \frac{12}{p-1} \frac{1+4\beta}{(1+2\beta)^2}.$$

If we denote by  $d[\beta_1, \beta_2]$  a standardised Pearsonian variable with  $\beta$ -ratios  $\beta_1$  and  $\beta_2$ , then the probability that  $F(p)$  exceeds a given quantity  $k$  is the same as greater than 1

$$P_r \left\{ d[\beta_1(p), \beta_2(p)] > \frac{k-1}{\sqrt{\frac{2}{p-1} \frac{1+2\beta}{(1+\beta)^2}}} \right\} \quad \dots (3.12)$$

and this is less than

$$P_r \left\{ d[\beta_1(p), \beta_2(p)] > \frac{k-1}{\sqrt{\frac{2}{p-1}}} \right\} \quad \dots (3.13)$$

Now it is seen from Pearson and Merrington's table (1951) that the upper tail area beyond a certain point in the range of distribution of a standardised Pearsonian variable is an increasing function of the  $\beta$ -coefficients individually.

Using this empirical result to hold true generally, we have

$$\rho_r \left\{ d[\beta_1(p), \beta_2(k)] > \frac{k-1}{\sqrt{\frac{2}{p-1}}} \right\} < P_r \left\{ d[\beta_1^*(p), \beta_2^*(p)] > \frac{k-1}{\sqrt{\frac{2}{p-1}}} \right\} \dots (3.14)$$

where

$$\beta_1^*(p) = \frac{8}{p-1} \quad \dots (3.15)$$

$$\beta_2^*(p) = \frac{12}{p-1} + 3$$

which are the  $\beta$ -coefficients of  $\frac{\chi_{p-1}^2}{p-1}$  where  $\chi_{p-1}^2$  is a central chi-square having  $p-1$  degrees of freedom, which are individually greater than  $\beta_1(p)$  and  $\beta_2(p)$  respectively.

Let now  $k$  be chosen such that

$$P_r \left\{ d[\beta_1^*(p), \beta_2^*(p)] > \frac{k-1}{\sqrt{\frac{2}{p-1}}} \right\} = \alpha\% \quad \dots (3.16)$$

i.e., let  $k$  be the upper  $\alpha\%$  point of the central  $F$ -distribution with  $p-1$  and  $\infty$  degrees of freedom. It follows from the inequality (3.14) that the probability of  $F(p)$  exceeding the  $\alpha\%$  curve in the  $F$ -diagram is less than  $\alpha\%$ .

#### 4. LIMITING DISTRIBUTION OF $F(p)$ WHEN THE SERIES IS LINEAR REGRESSIVE

4.1. We shall now assume that the observed series is Linear Regressive and normal and that  $\sum_{v=1}^{\infty} |\rho_v|$  converges. The density of  $F(p)$  then varies as the density of the ratio of two sums of squares of non-independent normal variables. We shall assume that  $N = np$ , that is, there are no fractional rows. This assumption does not in any way affect our conclusions.

4.2. Let  $Z_i^N = \{\bar{y}_{i\cdot}(p) - \bar{y}_{\cdot\cdot}\} \sqrt{n}$ , so that  $\sum_{i=1}^p \{\bar{y}_{i\cdot}(p) - \bar{y}_{\cdot\cdot}\}^2 n = \sum_{i=1}^p Z_i^{N^2}$ . The  $Z_i^N$  ( $i = 1, 2, \dots, p$ ) are distributed as multivariate normal variables with zero mean and a covariance matrix which we shall designate by  $[\sigma_{ij}(N)]$ . It can be shown that as  $N$  tends to infinity,  $\sigma_{ij}(N)$  and  $\sigma_{ii}(N)$  tend for all values of  $i$  and  $j$  to

$$\sigma_{ij} = V_0 \left\{ \rho_s - \frac{1}{p} + 2 \left( \tilde{\rho}(p) - \frac{\tilde{\rho}}{p} \right) \right\} \quad \dots (4.1)$$

and

$$\sigma_{ii} = V_0 \left\{ 1 - \frac{1}{p} + 2 \left( \tilde{\rho}(p) - \frac{\tilde{\rho}}{p} \right) \right\}. \quad \dots (4.2)$$

Where

$$s = |i-j|.$$

$$\tilde{\rho} = \lim_{N \rightarrow \infty} \sum_{i=1}^{N-1} \rho_i \quad \dots (4.3)$$

$$\tilde{\rho}(p) = \lim_{N \rightarrow \infty} \sum_{i=1}^{n-1} \rho_{ip} \quad \dots (4.4)$$

$$\tilde{\rho}(p) = \lim_{N \rightarrow \infty} \sum_{i=1}^{n-1} \left( \frac{\rho_{s+ip} + \rho_{s-ip}}{2} \right). \quad \dots (4.5)$$

The limits exist as a consequence of the existence of the limit  $\sum_{i=1}^{\infty} |\rho_i|$ .

4.3. The distribution of  $\sum_{i=1}^p Z_i^{N^2}$  is the same as that of  $\sum_{i=1}^p \lambda_i^2(N) \chi_i^2$  where  $\chi_i^2$  ( $i = 1, 2, \dots, p$ ) are independent chi-squares with one degree of freedom and

$\lambda_i(N)$  ( $i=1, 2, \dots, p$ ) the latent roots of the matrix  $[\sigma_{ij}(N)]$ , and its characteristic function is given by  $\prod_{i=1}^p \{1 - 2it\lambda_i(N)\}^{-\frac{1}{2}}$ . For any  $t$  this converges to  $\prod_{i=1}^p \{1 - 2it\lambda_i\}^{-\frac{1}{2}}$  as  $N \rightarrow \infty$ , where  $\lambda_i$  ( $i=1, 2, \dots, p$ ) are the latent roots of the matrix  $[\sigma_{ij}]$ ,  $\sigma_{ij}$  being given by (4.1) and (4.2). The distribution function of  $\sum_{i=1}^p Z_i^2$  therefore converges, as  $N$  tends to infinity to that of  $\sum_{i=1}^p \lambda_i \chi_i^2$ .

4.4. Let  $y_t = \{y_{ij}(p) - \bar{y}_i(p)\}$  where  $t$  is to be obtained from  $i$  and  $j$  by using the formula  $t = (i-1)p + j$ . Then  $y_t$  is a normal variable with mean zero and variance:

$$\sigma^2(y_t) = V_0 \left\{ 1 - \frac{1}{n} + \frac{n-1}{n} \bar{\rho}(n, p) - \frac{2}{n} \tilde{\rho}(u, p) - \tilde{\rho}(v, p) \right\} \quad \dots (4.6)$$

where 
$$(n-1)\bar{\rho}(n, p) = \frac{2}{n} \sum_{i=1}^{n-1} (n-i)\rho_{ip},$$

a function that converges to  $2\tilde{\rho}(p)$  as  $N \rightarrow \infty$ ;

$$\tilde{\rho}(w, p) = \rho_p + \rho_{2p} + \dots + \rho_{wp};$$

and  $u$  and  $v$  the smaller and larger of the values of  $n-j$  and  $j-1$ . As  $\sum_{r=1}^{\infty} |\rho_r|$  converges the sequence  $\tilde{\rho}(w, p)$  ( $w = 1, 2, \dots$ ) is bounded and lies between  $\pm \sum_{r=1}^{\infty} |\rho_r| = \pm C$  say.

Then 
$$\sigma^2(y_t) \leq V_0 \left\{ 1 - \frac{1}{n} + \frac{5}{n} C \right\}. \quad \dots (4.7)$$

It can be shown similarly that the correlation  $\rho(y_t y_{t'})$  between  $y_t$  and  $y_{t'}$  satisfy the inequality

$$|\rho(y_t y_{t'})| \leq \left\{ |\rho_\alpha| - \frac{1}{n} + \frac{5}{n} C \right\} \quad \dots (4.8)$$

when  $\alpha = |t - t'|$ .

Making use of the assumption of normality of  $y_t$  ( $t = 1, 2, \dots, N$ ), we now have for the second moment of  $\frac{1}{n-p} \sum_{ij} \{y_{ij}(p) - \bar{y}_i(p)\}^2$ :

$$\begin{aligned} \frac{2}{(N-p)^2} \left\{ \sum_{i=1}^N \sigma_i^2 + 2 \sum_{i \neq i'} \sigma_i^2 \sigma_{i'}^2 \rho^2(y_i y_{i'}) \right\} &\leq \frac{2V_0^2}{(N-p)^2} \left\{ 1 - \frac{1}{n} + \frac{5}{n} C \right\}^2 \times \\ &\times \left\{ N + 2 \sum_{s=1}^{N-1} (N-s) \left[ |\rho_s| - \frac{1}{n} + \frac{5}{n} C \right]^2 \right\}. \quad \dots (4.9) \end{aligned}$$



The limit of the right hand side as  $N \rightarrow \infty$  is equal to zero. As the second moment considered cannot be negative it follows that

$$\lim_{N \rightarrow \infty} \mu_2 \left\{ \sum_{ij} \frac{\{y_{ij}(p) - \bar{y}_{i\cdot}(p)\}^2}{N-p} \right\} = 0. \quad \dots (4.10)$$

4.5. The first moment of  $\frac{1}{N-p} \sum_{ij} \{y_{ij}(p) - \bar{y}_{i\cdot}(p)\}^2$  is of course  $\frac{1}{N-p} \sum_{i=1}^N \sigma_i^2$  and this is directly found to be  $V_0\{1 - \bar{p}(n, \rho)\}$ .

This tends to the limit  $V_0$ , as

$$\lim_{N \rightarrow \infty} \bar{p}(n, p) = \lim_{N \rightarrow \infty} \frac{2}{n(n-1)} \sum_{i=1}^{N-1} (n-i) \rho_{ip} = \lim_{N \rightarrow \infty} \frac{2}{n} \sum_{n-1}^{n-i} \rho_{ip} = 0. \quad \dots (4.11)$$

The statistic  $\frac{1}{N-p} \sum_{ij} \{y_{ij}(p) - \bar{y}_{i\cdot}(p)\}^2$  therefore converges in probability to  $V_0$ . The distribution function of the ratio  $F(p)$  therefore converges to that of

$$\frac{1}{p-1} \sum_{i=1}^p \frac{\lambda_i \chi_i^2}{V_0} \quad \dots (4.12)$$

the first two moments of which are

$$\begin{aligned} \mu_1\{F(p)\} &= \frac{1}{p-1} \sum_{i=1}^p \frac{\lambda_i}{V_0} = \frac{\text{tr}[\sigma_{ij}]}{(p-1)V_0} = \frac{1}{(p-1)V_0} \sum_{i=1}^p \sigma_{ii} \\ &= 1 + \frac{2p}{p-1} \left\{ \tilde{\rho}(p) - \frac{\tilde{\rho}}{p} \right\} \quad \dots (4.13) \end{aligned}$$

$$\begin{aligned} \text{and } \mu_2\{F(p)\} &= \frac{1}{(p-1)^2 V_0^2} \left[ \left\{ 3 \sum_{i=1}^p \lambda_i^2 + 2 \sum_{i \neq j} \lambda_i \lambda_j \right\} - \left\{ \sum_{i=1}^p \lambda_i \right\}^2 \right] \\ &= \frac{2}{(p-1)^2 V_0^2} \left\{ \sum_{i=1}^p \lambda_i^2 \right\} \quad \dots (4.14) \end{aligned}$$

$$\begin{aligned}
 &= \frac{2}{(p-1)^2 V_0^2} \text{tr} [\sigma_{ij}]^2 = \frac{2}{(p-1)^2 V_0^2} \left\{ \sum_{i=1}^p \sigma_{ii}^2 + 2 \sum_{i \neq j} \sigma_{ij}^2 \right\} \\
 &= \frac{2}{(p-1)^2} \left\{ 1 - \frac{1}{p} + 2 \left( \tilde{\rho}(p) - \frac{\tilde{\rho}}{p} \right) \right\}^2 \left\{ p + 2 \sum_{i \neq j} \rho_{ij}^2 \right\} \sum_{i \neq j} \rho_{ij}
 \end{aligned}$$

$$\text{where } \rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}^2} \sqrt{\sigma_{jj}^2}} = \frac{\rho_{ij} - \frac{1}{p} + 2 \left( \tilde{\rho}(p) - \frac{\tilde{\rho}}{p} \right)}{1 - \frac{1}{p} + 2 \left( \tilde{\rho}(p) - \frac{\tilde{\rho}}{p} \right)} \quad \dots (4.15)$$

$$\simeq \frac{2}{p} + \frac{8}{p-1} \left\{ \tilde{\rho}(p) - \frac{\tilde{\rho}}{n} \right\} + \frac{4}{p} \sum_{s=1}^{p-1} \frac{(p-s)}{p} \rho_s^2 \quad \dots (4.16)$$

4.6.  $\rho_1 \rho_2 \dots \rho_{N-1}$  being the autocorrelations of a Stationary Process, they must satisfy the condition

$$\sum_{ij}^N x_i x_j \rho_{|i-j|} > 0$$

for any set of real numbers  $x_i$  ( $i = 1, 2, \dots, N$ ).

Let  $x_1 = x_2 = \dots = x_n = 1$ . Then,

$$N + 2 \sum_{s=1}^{N-1} (N-s) \rho_s > 0$$

$$\text{or} \quad N + N(N-1) \bar{\rho}(N) > 0$$

$$\text{where} \quad \bar{\rho}(N) = \frac{2}{N(N-1)} \sum_{s=1}^{N-1} (N-s) \rho_s$$

$$\text{or} \quad \bar{\rho}(N) > -\frac{1}{N-1} \quad \dots (4.17)$$

Now, for not too small  $p$ ,

$$\mathcal{E}[F(p)] \simeq 1 - n \bar{\rho}(N) \quad \dots (4.18)$$

so that 
$$\mathcal{E}[F(p)] \leq 1 + \frac{n}{N-1} \simeq 1 + \frac{1}{p}. \quad \dots (4.19)$$

The first moment of  $F(p)$  therefore cannot exceed unity by any large quantity.

Multiplying the inequality (4.17) by  $N$  and proceeding to limit on both sides, we have

$$\sum_{i=1}^{\infty} \rho_i = \tilde{\rho} > -\frac{1}{2}. \quad \dots (4.20)$$

It is interesting to obtain expressions for  $\tilde{\rho}$  for particular cases. Let the series be a moving average:

$$x_t = \sum_{i=0}^k \alpha_i e_{t-i}, \quad (t = 1, 2, \dots, N). \quad \dots (4.21)$$

Then, 
$$\rho_s = \frac{\alpha_0 \alpha_s + \alpha_1 \alpha_{s+1} + \dots + \alpha_{k-s} \alpha_k}{\alpha_0^2 + \alpha_1^2 + \dots + \alpha_k^2} \quad \dots (4.22)$$

and 
$$\tilde{\rho} = \sum_{s=1}^{\infty} \rho_s = \frac{1}{2} \left\{ \frac{(\alpha_0 + \alpha_1 + \dots + \alpha_k)^2}{\alpha_0^2 + \alpha_1^2 + \dots + \alpha_k^2} - 1 \right\}. \quad \dots (4.23)$$

$\tilde{\rho}$  therefore lies between  $\frac{K}{2}$  and  $-\frac{1}{2}$ .

Let us now consider an autoregressive process of order  $k$  having the characteristic roots  $\gamma, \gamma_2 \dots \gamma_k$ . Then

$$\rho_s = \sum_{i=1}^k A_i \gamma_i^s \quad \dots (4.24)$$

where  $A_i$  ( $i = 1, 2, \dots, k$ ) are constants.

Hence 
$$\tilde{\rho} = \sum_{i=1}^k A_i \frac{\gamma_i}{1-\gamma_i}, \text{ which however has no upper bound.}$$

## 5. THE COMPUTATIONAL ASPECT OF THE METHOD

5.1. An important point in favour of the  $F$ -diagram method as against the previous method of Schuster's is that it takes much less computational labour. For the present method involves only the squaring of a large number of figures which can be done by a computing machine reasonably quickly. Schuster's methods involves the calculation of the covariance between the group totals in the Buys Ballot Table and certain sine and cosine terms which vary for every period length and which require the use of appropriate tables. The calculation of  $F(p)$  involves only the calculation of (1) the total sum of squares  $\sum_{t=1}^N x_t^2 - \bar{x}^2 \cdot N$  and (2) the between group sum of squares  $\sum_{i=1}^p \bar{y}_{i.}^2 (p)n(p) - N\bar{y}_{..}^2$ .



## A METHOD OF DISCRIMINATION IN TIME SERIES ANALYSIS—II

Hence if our diagram is to consist of  $k$  ordinates, we need calculate only  $k+1$  sums of squares which can be done reasonably quickly. For a long series or when there are several series to analyse, means may be devised to simplify the work. In analysis of isolated series, the observations may be entered on cards and the cards dealt into  $p$  groups, where  $p$  is the trial period. The total of each group will then give the corresponding column total of the Buys Ballot Table. It was however found that for a not-too-long series, it would be quicker to write down the observations in actual Buys Ballot arrangements and get the column totals directly from these.

When analysis is to be carried out on a routine basis, the punched-card-method can be used with advantage to get the column totals of the Buys Ballot Table. We have actually made use of Hollerith equipment in simultaneously analysing 25 series.

5.2. Schuster's method of Periodogram analysis can be directly applied to integral as well as rational fraction periodicities though the ordinates for periods of irrational length can only be interpolated. The  $F$ -diagram method can however be directly applied only to integral periodicities as the method is based on the arrangement of the series in a Buys Ballot Table, and a Buys Ballot Table with a fractional number of columns is not possible. The present method can however be adapted to the purpose of fractional periodicities by various means. Two methods are described below. The first, used by Whittaker, for his own periodogram and described in the book "Calculus of Observations", is useful only in regard to fractions that can be expressed as  $p + \frac{1}{M}$  where  $p$  and  $M$  are positive integers,  $M$  preferably small. The second method is useful for any fractional length that can be measured in a scale.

A. The method for a trial period of length  $(p + \frac{1}{M})$  where  $M$  is a positive integer, is to arrange the data in a table which would be a Buys Ballot Table for trial period  $p$  but for the fact that a  $(p+1)$ th column is formed by the elements  $x_{pk+1}$ ,  $x_{2pk+1}$ ,  $x_{3pk+1}$  etc., so as to bring the elements of the first, the  $(k+1)$ th the  $(2k+1)$ th etc., rows in phase, and similarly with every pair of rows with a gap of  $k$  rows in between them. Thus, for trial period  $5\frac{1}{2}$  the data should be arranged as follows:

Columns:	(1)	(2)	(3)	(4)	(5)	(6)
	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_{11}$
	$x_6$	$x_7$	$x_8$	$x_9$	$x_{10}$	
	$x_{12}$	$x_{13}$	$x_{14}$	$x_{15}$	$x_{16}$	$x_{22}$
	$x_{17}$	$x_{18}$	$x_{19}$	$x_{20}$	$x_{21}$	
	$x_{23}$	...				

For trial period  $4\frac{1}{2}$ , the data should be arranged as follows:

Columns:	(1)	(2)	(3)	(4)	(5)
	$x_1$	$x_2$	$x_3$	$x_4$	
	$x_5$	$x_6$	$x_7$	$x_8$	
	$x_9$	$x_{10}$	$x_{11}$	$x_{12}$	$x_{13}$
	$x_{14}$	$x_{15}$	...		

The argument is that the concept of fractional periodicity is acceptable only on the basis of an underlying continuous stochastic phenomenon. If we assume that the amount of variation between possible values between any two successive observations is much less than the total variations among the observations in a complete cycle, the confounding of different phrases in the same column will not have any very serious effect. Hence the  $F$ -ratio of the between-column variation of the  $(p+1)$  columns to the within column variation should serve as a test criterion for the periodicity  $\left(p + \frac{1}{M}\right)$  for reasons exactly similar to those in the case of integral trial periods. It should be noted that for all fractional trial periods between  $p$  and  $p+1$ , the degrees of freedom of  $F(p)$  are  $p$  and  $N-p-1$ . Hence, the significance curve for the  $F$ -diagram will not be continuous but will be step like.

B. The second method consists in dividing up the time continuum from 1 to  $N$  into intervals of length  $\frac{p+\delta}{p}$  where  $p+\delta$  is the trial period,  $p$  being a positive integer and  $\delta$  any positive proper fraction. Each interval will contain at least one and at most two observations. Let the intervals be marked 1, 2, ... etc., up to  $(r)$  where  $N = \frac{p+\delta}{p} r + s$  and  $s < \frac{p+\delta}{p}$ . Let the intervals, each of which carry one or at most two observations, be now arranged in a Buys Ballot Table with  $p$  columns. This can be done by writing up the series on a tape, the entire being at equidistant points, cutting up the tape into lengths of  $p+\delta$  units, making out each piece into  $p$  compartments, and arranging the pieces in succession one below the other, and adding up all the observations that belong to compartments in the same column.

This method involves the use of  $(p-1)$  and  $N-p+1$  degrees of freedom for all fractional periods between  $p$  and  $p+1$ . The significance curve in the  $F$ -diagram is again broken, and is everywhere one step above the significance curve of the first method.

While the first method groups together in the same column, observations which differ in phase by less than the phase difference between two successive observations, the second method by actually grouping together in the same column two successive observations confounds phase differences actually equalling the phase difference between two successive observations. Hence the second method is less accurate; it is also more elaborate; but it does have the advantage of being useful for any fraction while the first method is useful only for such values of  $\delta$  as  $\frac{1}{2}, \frac{1}{3}, \frac{1}{4}$ ...etc.

#### REFERENCES

- PATNAIK, P. B. (1949): The non-central  $\chi^2$  and  $F$ -distributions and their applications. *Biometrika*, **36**, 202.
- PEARSON, E. S. and MERRINGTON, M. (1951): Tables of the 5% and 0.5% points of the Pearson Curves. *Biometrika*, **38**, 4.
- RUDRA, A. (1955): A method of discrimination in time series analysis-I. *Sankhyā*, **15**, 9-34.
- WHITTAKER, E. T. and ROBINSON, G. (1940): The Calculus of Observations, Blackies and Sons, London.
- Paper received : June, 1954, revised April, 1955.*

# ON THE TESTING OF OUTLYING OBSERVATIONS

By A. KUDO

*Kyusyu University, Fukuoka*  
and  
*Indian Statistical Institute, Calcutta*

## 1. INTRODUCTION

The problem of testing outlying observations concerning the mean value of normal population has been treated by various authors such as W. R. Thompson (1942), E. S. Pearson and C. Chandrasekar (1936), K. R. Nair (1948), F. E. Grubbs (1950) and N. V. Smirnov (1940). The statistics suggested by these authors are mostly based on the difference between the extreme value and the mean value, when the variances are assumed to be same and known to us. When the variance is unknown, the statistic is the difference divided by the estimate of the standard deviation. Concerning the latter case the statistics treated by Grubbs and Smirnov seem to be essentially the same where the standard deviation is estimated from the sample in hand, whereas K. R. Nair suggested that the standard deviation should be estimated from another independent sample. The former is called the Pearson-Chandrasekar statistic. The characters of these statistics, like efficiency or optimum property of power function, have not been fully treated so far. In this paper, we propose statistics to meet the more general situations, where we have got other independent samples containing information about the population mean value and the variance of the population. The Pearson-Chandrasekar statistic is a special case of ours. We shall prove that our statistic is optimum in the sense that it is uniformly best for special class of alternative hypotheses among a certain reasonably restricted class of testing procedures.

In § 2, we shall formulate the problem in the case when the variance is unknown and we shall propose a testing procedure, which will be proved to be optimum in § 3. As a corollary, we shall observe that the Pearson-Chandrasekar statistic is optimum whereas Nair's statistic is not. In § 4, we shall discuss briefly the problem in the case when the variance is known to us.

Concerning the tables to be used for our test procedure, some results have been obtained by the author, which will be discussed in a forthcoming paper.

## 2. FORMULATION OF THE PROBLEM

Let  $x_i^{(1)}$  ( $i = 1, 2, \dots, N_1$ ) be distributed as  $N(m_i, \sigma^2)$  ( $i = 1, 2, \dots, N_1$ ) respectively, and  $x_i^{(2)}$  ( $i = 1, 2, \dots, N_2$ ) as  $N(m^{(2)}, \sigma^2)$  and  $x_i^{(3)}$  ( $i = 1, 2, \dots, N_3$ ) as  $N(m^{(3)}, \sigma^2)$ . We assume that they are mutually independent and the values of these parameters  $m_i$  ( $i = 1, 2, \dots, N_1$ ),  $m^{(2)}$ ,  $m^{(3)}$  and  $\sigma$  are unknown to us.



Our null hypothesis  $H_0$  is  $H_0 = H(m_1 = m_2 = \dots = m_{N_1} = m^{(2)})$  where  $m^{(2)}$ ,  $m^{(3)}$  and  $\sigma$  are free. We have  $N_1$  alternative hypotheses  $H_1, H_2, \dots, H_{N_1}$  where  $H_i (i = 1, 2, \dots, N_1)$  is the hypothesis  $H_i = H(m_1 = \dots = m_{i-1} = m_i - \Delta = m_{i+1} = \dots = m_{N_1} = m^{(2)}, \Delta > 0)$ . In other words, under  $H_0$  the observations in the first and the second groups are all from the same population, while under the hypothesis  $H_i$  the  $i$ -th observation in the first group only is from a different normal population with greater mean and the same variance.

Let  $D_i (i = 0, 1, 2, \dots, N_1)$  be the decision to accept  $H_i (i = 0, 1, 2, \dots, N_1)$ . Our problem is to find out an optimum decision procedure as to these  $N_1 + 1$  decisions.

At first, we have to make the meaning of optimum clear. We introduce the following requirements or criteria:

(1<sup>0</sup>) We want to accept the decision  $D_0$  with the pre-assigned probability, say,  $1-p$ , when  $D_0$  is true.

(2<sup>0</sup>) The decision procedure must be invariant (a) the addition of any constant to all the first  $N_1 + N_2$  observations and (b) under the addition of any constant to all the last  $N_3$  observations.

(3<sup>0</sup>) The decision procedure must remain invariant when all the observations are multiplied by any positive constant.

The last two conditions require that if we have two groups of observations  $(x_1^{(1)}, \dots, x_{N_1}^{(1)}, x_1^{(2)}, \dots, x_{N_2}^{(2)}, x_1^{(3)}, \dots, x_{N_3}^{(3)})$  and  $(y_1^{(1)}, \dots, y_{N_1}^{(1)}, y_1^{(2)}, \dots, y_{N_2}^{(2)}, y_1^{(3)}, \dots, y_{N_3}^{(3)})$  and there are the following relations

$$y_j^{(i)} = ax_j^{(i)} + b, \quad (i = 1, 2; j = 1, 2, \dots, N_i)$$

and

$$y_j^{(3)} = ax_j^{(3)} + c, \quad (j = 1, 2, \dots, N_3),$$

where  $a$  is positive and  $b$  and  $c$  are constant, our decision procedure must give the same decision.

(4<sup>0</sup>) The probability of a correct decision when the  $i$ -th population mean is shifted to the right by the same amount must be the same for all  $i$ .

(5<sup>0</sup>) We want to maximize the probability of making a correct decision when  $D_i$  is correct.

Now our problem is to find out a decision procedure satisfying the conditions (1<sup>0</sup>)—(5<sup>0</sup>).

Let 
$$\bar{x}^{(i)} = \sum_{j=1}^{N_i} x_j^{(i)} / N_i, \quad \bar{x} = (N_1 \bar{x}^{(1)} + N_2 \bar{x}^{(2)}) / (N_1 + N_2)$$

$$\bar{m}^{(1)} = \sum_{i=1}^{N_1} \bar{m}_i / N_1, \quad \bar{m} = (N_1 \bar{m}^{(1)} + N_2 m^{(2)}) / (N_1 + N_2),$$

# ON THE TESTING OF OUTLYING OBSERVATIONS

$$\left. \begin{aligned}
 S_i^2 &= \sum_{j=1}^{N_i} (x_j^{(i)} - \bar{x}^{(i)})^2 / N_i \\
 S_{12}^2 &= \left\{ \sum_{j=1}^{N_1} (x_j^{(1)} - \bar{x})^2 + \sum_{j=1}^{N_2} (x_j^{(2)} - \bar{x})^2 \right\} / (N_1 + N_2) \\
 S^2 &= \{(N_1 + N_2)S^2 + N_3 S_3^2\} / (N_1 + N_2 + N_3) \\
 S_m^2 &= \left\{ \sum_{j=1}^{N_1} (m_j^{(1)} - \bar{m})^2 + N_2 (m^{(2)} - \bar{m})^2 \right\} / (N_1 + N_2) \\
 S_m^2(a) &= \frac{(N_1 + N_2 - 1)}{(N_1 + N_2)^2} a^2 \\
 (i &= 1, 2, 3).
 \end{aligned} \right\} \dots (2.1)$$

Let  $M$  be the suffix of the population which has the greatest sample value among the first  $N_1$  observations, so that

$$x_M = \max_{j=1, 2, \dots, N_1} x_j^{(1)}.$$

Our optimum decision procedure will be shown to be as follows:

$$\left. \begin{aligned}
 \text{if } \frac{x_M - \bar{x}}{S} &\leq \lambda_p, \text{ select } D_0, \\
 \text{if } \frac{x_M - \bar{x}}{S} &> \lambda_p, \text{ select } D_M,
 \end{aligned} \right\} \dots (2.2)$$

where  $\lambda_p$  is a constant which depends on  $N_1, N_2, N_3$  and  $p$ . Concerning the numerical calculation of  $\lambda_p$ , when  $N_2 = N_3 = 0$ , the values have been tabulated by ENIAC (see Grubbs, 1950).

## 3. DERIVATION OF THE OPTIMUM PROCEDURE

In the usual manner, our decision procedure can be expressed by the following  $N_1 + 1$  functions,  $d_0(\tilde{X}), d_1(\tilde{X}), \dots, d_{N_1}(\tilde{X})$  where  $\tilde{X}$  denotes the observation  $(x_1^{(1)}, \dots, x_{N_1+1}^{(1)}, x_1^{(2)}, \dots, x_{N_2}^{(2)}, x_1^{(3)}, \dots, x_{N_3}^{(3)})$  and  $d_i(\tilde{X}), (i = 0, 1, \dots, N_1)$  denote the probabilities of selecting  $D_i, (i = 0, 1, \dots, N_1)$  respectively when the observation  $\tilde{X}$  is given. We shall call this set of  $N_1 + 1$  functions a decision function. Naturally these  $N_1 + 1$  functions must satisfy the following relations  $1 \geq d_i(\tilde{X}) \geq 0, (i = 0, 1, \dots, N_1)$  and  $\sum_{i=0}^{N_1} d_i(\tilde{X}) = 1$ . To meet our mathematical necessity we shall assume that they are measurable functions.

By condition (2<sup>o</sup>) we see that our decision procedure must depend only on the  $N_1 + N_2 + N_3 - 2$  values of

$$\left. \begin{aligned}
 y_1^{(1)} &= x_1^{(1)} - x_{N_1}^{(1)}, \dots, y_{N_1-1}^{(1)} = x_{N_1-1}^{(1)} - x_{N_1}^{(1)} \\
 y_1^{(2)} &= x_1^{(2)} - x_{N_1}^{(1)}, \dots, y_{N_2}^{(2)} = x_{N_2}^{(2)} - x_{N_1}^{(2)} \\
 y_1^{(3)} &= x_1^{(3)} - x_{N_3}^{(3)}, \dots, y_{N_3-1}^{(3)} = x_{N_3-1}^{(3)} - x_{N_3}^{(3)}.
 \end{aligned} \right\} \dots (3.1)$$

Further by condition (3<sup>0</sup>) we see that our decision procedure must depend only on the  $N_1 + N_2 + N_3 - 2$  values of

$$\left. \begin{aligned} Z_1^{(1)} &= \frac{y_1^{(1)}}{|y_{N_1-1}^{(1)}|}, \dots, Z_{N_1-2}^{(1)} = \frac{y_{N_1-2}^{(1)}}{|y_{N_1-1}^{(1)}|}, Z_{N_1-1}^{(1)} = \frac{y_{N_1-1}^{(1)}}{|y_{N_1-1}^{(1)}|}, \\ Z_1^{(2)} &= \frac{y_1^{(2)}}{|y_{N_1-1}^{(1)}|}, \dots, Z_{N_2}^{(2)} = \frac{y_{N_2}^{(2)}}{|y_{N_1-1}^{(1)}|}, \\ Z_1^{(3)} &= \frac{y_1^{(3)}}{|y_{N_1-1}^{(1)}|}, \dots, Z_{N_3-1}^{(3)} = \frac{y_{N_3-1}^{(3)}}{|y_{N_1-1}^{(1)}|}. \end{aligned} \right\} \dots \quad (3.2)$$

Clearly the joint distribution of  $\tilde{Z}_j^{(i)}$ 's does not depend on any parameter of the population when  $D_0$  is true, while it depends on the value of  $\Delta/\sigma$  and the suffix  $i$  when  $D_i$  is correct.

Let  $\tilde{Z}$  denote  $N_1 + N_2 + N_3 - 2$  ( $= M$  say) sample values;  $f_0(\tilde{Z})$  be the frequency function of  $\tilde{Z}$  when  $D_0$  is correct;  $f_i(\tilde{Z}|a)$  be the frequency function of  $\tilde{Z}$  when  $D_i$  is correct; and  $a = \Delta/\sigma$ . As we have seen the decision function of our procedure must be a function of  $\tilde{Z}$  only, and hence we can write our decision function as  $d_0(\tilde{Z}), d_1(\tilde{Z}), \dots, d_{N_1}(\tilde{Z})$ .

Then our problem can be formulated as follows. We want to find the decision function  $(d_0(\tilde{Z}), d_1(\tilde{Z}), \dots, d_{N_1}(\tilde{Z}))$  such that

$$\int d_0(\tilde{Z}) f_0(\tilde{Z}) d\tilde{Z} = 1 - p \quad \dots \quad (3.3)$$

$$\text{and} \quad \int d_i(\tilde{Z}) f_i(\tilde{Z}|a) d\tilde{Z} \quad \dots \quad (3.4)$$

is independent of  $i$  and is maximum.

Here we need the frequency function of  $\tilde{Z}$ . At first we shall find the probability density function of  $y_j^{(i)}$ 's. As  $f_0(\tilde{Z})$  does not depend on any parameter and  $f_i(\tilde{Z}|a)$  depends only on  $a$ , we can assume that  $\sigma = 1$ ,  $m^{(2)} = 0$ ,  $m^{(3)} = 0$ .

$$\text{Let} \quad M_1 = m_1 - m_{N_1}, \dots, M_{N_1-1} = m_{N_1-1} - m_{N_1}. \quad \dots \quad (3.5)$$

Then  $y_1^{(1)}, \dots, y_{N_1-1}^{(1)}, y_1^{(2)}, \dots, y_{N_2}^{(2)}$  are normally distributed with mean values  $M_1, \dots, M_{N_1-1}, -m_{N_1}, \dots, -m_{N_1}$  and common variances equal to 2 and common covariances equal to 1; while  $y_1^{(3)}, \dots, y_{N_3-1}^{(3)}$  are distributed with mean values equal to zero and with



# ON THE TESTING OF OUTLYING OBSERVATIONS

same variances and covariances as those of the former group. and these two groups are independent. Therefore the probability density function is given by

$$C \exp \left[ -\frac{1}{2} \frac{1}{N_1 + N_2} \left\{ (N_1 + N_2) \left( \sum_{a=1}^{N_1-1} (y_a^{(1)} - M_a)^2 + \sum_{a=1}^{N_2} (y_a^{(2)} + m_{N_1})^2 \right) - \right. \right. \\ \left. \left. - \left( \sum_{a=1}^{N_1-1} (y_a^{(1)} - M_a) + \sum_{a=1}^{N_2} (y_a^{(2)} + m_{N_1}) \right)^2 \right\} \right] \times \\ \times \exp \left[ -\frac{1}{2} \frac{1}{N_3} \left\{ N_3 \sum_{a=1}^{N_3-1} y_a^{(3)2} - \left( \sum_{a=1}^{N_3-1} y_a^{(3)} \right)^2 \right\} \right] \dots (3.6)$$

where  $C$  is some constant the exact value of which is not needed for our purpose.

By a simple transformation we get the frequency function of

$$f(\tilde{Z}) = C P_r(y_{N_1-1}^{(1)} > 0) \int_0^\infty t^M \exp \left[ -\frac{1}{2} \frac{1}{N_1 + N_2} \left\{ (N_1 + N_2) \left\{ \sum_{a=1}^{N_1-2} (Z_a^{(1)} t - M_a)^2 + (t - M_{N_1-1})^2 + \right. \right. \right. \\ \left. \left. + \sum_{a=1}^{N_2} (Z_a^{(2)} t + m_{N_1})^2 \right\} - \left( \sum_{a=1}^{N_1-2} (Z_a^{(1)} t - M_a) + (t - M_{N_1-1}) + \sum_{a=1}^{N_2} (Z_a^{(2)} t + m_{N_1}) \right)^2 \right\} \right] \times \\ \times \exp \left[ -\frac{1}{2} \frac{t^2}{N_3} \left\{ N_3 \sum_{a=1}^{N_3-1} Z_a^{(3)2} - \left( \sum_{a=1}^{N_3-1} Z_a^{(3)} \right)^2 \right\} \right] dt + \\ + C P_r(y_{N_1-1}^{(1)} < 0) \int_0^\infty t^M \exp \left[ -\frac{1}{2} \frac{1}{N_1 + N_2} \left\{ (N_1 + N_2) \left\{ \sum_{a=1}^{N_1-2} (-Z_a^{(1)} t - M_a)^2 + (-t - M_{N_1-1})^2 + \right. \right. \right. \\ \left. \left. + \sum_{a=1}^{N_2} (-Z_a^{(2)} t + m_{N_1})^2 \right\} - \left( \sum_{a=1}^{N_1-2} (-Z_a^{(1)} t - M_a) - (-t - M_{N_1-1}) + \sum_{a=1}^{N_2} (-Z_a^{(2)} t + m_{N_1}) \right)^2 \right\} \right] \times \\ \times \exp \left[ -\frac{1}{2} \frac{t^2}{N_3} \left\{ N_3 \sum_{a=1}^{N_3-1} Z_a^{(3)2} - \left( \sum_{a=1}^{N_3-1} Z_a^{(3)} \right)^2 \right\} \right] dt. \dots (3.7)$$

As we have relations (3.1), (3.2) and (3.5) we can easily see that this is equal to

$$C P_r \left( Z_{N_1-1} = \frac{y_{N_1-1}^{(1)}}{|y_{N_1-1}^{(1)}|} \right) \int_0^\infty t^M \exp \left[ -\frac{1}{2} \frac{1}{N_1 + N_2} \left\{ (N_1 + N_2) \left\{ \sum_{a=1}^{N_1-2} \left( \frac{y_a^{(1)}}{y_{N_1-1}^{(1)}} t - M_a \right)^2 + \right. \right. \right. \\ \left. \left. + (t - M_{N_1-1})^2 + \sum_{a=1}^{N_2} \left( \frac{y_a^{(2)}}{y_{N_1-1}^{(1)}} t + m_{N_1} \right)^2 \right\} - \right. \\ \left. - \left( \sum_{a=1}^{N_1-2} \left( \frac{y_a^{(1)}}{y_{N_1-1}^{(1)}} \right) t - M_a \right) + (t - M_{N_1-1}) + \sum_{a=1}^{N_2} \left( \frac{y_a^{(2)}}{y_{N_1-1}^{(1)}} t + m_{N_1} \right) \right\} \right] \times$$

$$\begin{aligned}
& \times \exp \left[ -\frac{1}{2} \frac{t^2}{N_3} \left\{ N_3 \sum_{\alpha=1}^{N_3-1} \left( \frac{y_{\alpha}^{(3)}}{y_{N_1-1}^{(1)}} \right)^2 - \left( \sum_{\alpha=1}^{N_3-1} \frac{y_{\alpha}^{(3)}}{y_{N_1-1}^{(1)}} \right)^2 \right\} \right] dt \\
& = C P_r \left( Z_{N_1-1} = \frac{y_{N_1-1}^{(1)}}{|y_{N_1-1}^{(1)}|} \right) \int_0^{\infty} t^M \exp \left[ -\frac{1}{2} \frac{1}{N_1+N_2} \frac{1}{y_{N_1-1}^{(1)2}} \left[ (N_1+N_2) \times \right. \right. \\
& \quad \times \left\{ \sum_{\alpha=1}^{N_1-2} (y_{\alpha}^{(1)} t - y_{N_1-1}^{(1)} M_{\alpha})^2 + y_{N_1-1}^{(1)2} (t - M_{N_1-1})^2 + \sum_{\alpha=1}^{N_2} (y_{\alpha}^{(2)} t + y_{N_1-1}^{(1)} m_{N_1})^2 \right\} - \\
& \quad \left. - \left\{ \sum_{\alpha=1}^{N_1-1} y_{\alpha}^{(1)} t - y_{N_1-1}^{(1)} \left( \sum_{\alpha=1}^{N_1-1} M_{\alpha} \right) + \sum_{\alpha=1}^{N_2} y_{\alpha}^{(2)} t + y_{N_1-1}^{(1)} m_{N_1} \right\}^2 \right] \times \\
& \quad \times \exp \left[ -\frac{1}{2} \frac{t^2}{N_3} \frac{1}{y_{N_1-1}^{(1)2}} \left\{ N_3 \left( \sum_{\alpha=1}^{N_3-1} y_{\alpha}^{(3)2} - \left( \sum_{\alpha=1}^{N_3-1} y_{\alpha}^{(3)} \right)^2 \right) \right\} \right] dt \\
& = C |y_{N_1-1}^{(1)}|^{M+1} \int_0^{\infty} t^M \exp \left[ -\frac{1}{2} \frac{1}{N_1+N_2} \left[ t^2 \left\{ (N_1+N_2) \left( \sum_{\alpha=1}^{N_1-1} y_{\alpha}^{(1)2} + \sum_{\alpha=1}^{N_2} y_{\alpha}^{(2)2} \right) - \right. \right. \right. \\
& \quad \left. \left. \left. - \left( \sum_{\alpha=1}^{N_1-1} y_{\alpha}^{(1)} + \sum_{\alpha=1}^{N_2} y_{\alpha}^{(2)} \right)^2 \right\} - \right. \right. \\
& \quad \left. \left. - 2t \left\{ (N_1+N_2) \left( \sum_{\alpha=1}^{N_1-1} y_{\alpha}^{(1)} M_{\alpha} - \sum_{\alpha=1}^{N_2} y_{\alpha}^{(2)} m_{N_1} \right) - \left( \sum_{\alpha=1}^{N_1-1} y_{\alpha}^{(1)} + \sum_{\alpha=1}^{N_2} y_{\alpha}^{(2)} \right) \left( \sum_{\alpha=1}^{N_1-1} M_{\alpha} - N_2 m_{N_1} \right) \right\} + \right. \right. \\
& \quad \left. \left. + \left\{ (N_1+N_2) \left( \sum_{\alpha=1}^{N_1-1} M_{\alpha}^2 + N_2 m_{N_1}^2 \right) - \left( \sum_{\alpha=1}^{N_1-1} M_{\alpha} - N_2 m_{N_1} \right)^2 \right\} \right] \right] \times \\
& \quad \times \exp \left[ -\frac{1}{2} \frac{t^2}{N_3} \left\{ N_3 \left( \sum_{\alpha=1}^{N_3} y_{\alpha}^{(3)2} \right) - \left( \sum_{\alpha=1}^{N_3} y_{\alpha}^{(3)} \right)^2 \right\} \right] dt \\
& = C |x_{N_1-1}^{(1)} - x_{N_1}^{(1)}|^{M+1} \exp \left[ -\frac{1}{2} (N_1+N_2) S^2 \right] \times \\
& \quad \times \int_0^{\infty} \exp \left[ -\frac{t^2}{2} (N_1+N_2+N_3) S^2 \right] \exp \left[ t (N_1+N_2) \left\{ \sum_{\alpha=1}^{N_1} x_{\alpha}^{(1)} m_{\alpha} - \bar{x} \left( \sum_{\alpha=1}^{N_1} m_{\alpha} \right) \right\} \right] dt. \dots (3.8)
\end{aligned}$$

# ON THE TESTING OF OUTLYING OBSERVATIONS

It should be noted here that this is not the joint density function of  $\tilde{X}$ , but is only an expression for the frequency function of  $\tilde{Z}$  in terms of  $\tilde{X}$ . This enables us to make our discussion simple.

Now let us consider

$$\int A d_0(\tilde{Z}) f_0(\tilde{Z}) + \sum_{i=1}^{N_1} d_i(\tilde{Z}) f_i(\tilde{Z} | a) d\tilde{Z}. \quad \dots (3.9)$$

This value depends on the decision function and the values of  $A$  and  $a$ . The decision function for which (3.9) attains the maximum value when  $A$  and  $a$  fixed, is easily proved to be the following

$$d_0(\tilde{Z}) = 1 \quad \text{if } A f_0(\tilde{Z}) > f_j(\tilde{Z}) \text{ for all } j; \quad \dots (3.10)$$

$$d_i(\tilde{Z}) = 1 \quad \text{if } f_i(\tilde{Z}) > A f_0(\tilde{Z}), f_i(\tilde{Z}) > f_j(\tilde{Z}) \quad \dots (3.11)$$

for all  $j(i \neq j)$ .

For other points we define the value of decision function and arbitrarily, because it does not affect the value of the integral (3.9).

We shall prove that for any positive value  $a$  and probability  $p$ ,  $0 \leq p \leq 1$ , there exists a positive number  $A$  such that (3.9) attains its maximum for the decision function (2.2) with probability  $1-p$  of selecting  $D_0$  when  $D_0$  is the correct decision.

It will then immediately follow that the procedure (2.2) is optimum. Because if it is not optimum, there must be another decision procedure with the same value for the integral (3.3) but with a greater value for (3.4). Therefore, the value of (3.9) must increase. This leads us to a contradiction.

It is sufficient to prove that the following relations hold for some suitably chosen  $A$ .

$$A f_0(\tilde{Z}) \geq f_i(\tilde{Z} | a) \text{ if and only if } (x_i - \bar{x})/s \leq \lambda_p. \quad \dots (3.12)$$

$$f_i(\tilde{Z} | a) \geq f_j(\tilde{Z} | a) \text{ if and only if } x_i \geq x_j. \quad \dots (3.13)$$

This is so since these relations show that the decision function (3.2) makes the value of (3.9) maximum because of the relations (3.10) and (3.11). As

$$\{\tilde{X}; f_i(\tilde{Z} | a) \geq f_j(\tilde{Z} | a)\} = \{\tilde{X}; g_i(\tilde{X} | a) \geq g_j(\tilde{X} | a)\}$$

where  $g_i(\tilde{X} | a) = g(\tilde{X} | m_1 = \dots = m_{i-1} = m_{i+1} = \dots = m_{N_1} = 0, m_i = a)$ .

We have

$$g_i(\tilde{X} | a) - g_j(\tilde{X} | a) = C |x_{N_1}^{(1)} - 1 - x_{N_1}^{(1)}|^{M+1} \exp \left( -\frac{1}{2} (N_1 + N_2) S_m^2(a) \right) \times \\ \times \int_0^\infty t^M \exp \left( -\frac{t^2}{2} (N_1 + N_2 + N_3) S^2 \right) \exp \left[ \left( at(x_i - \bar{x}) - \exp(-at(x_j - \bar{x})) \right) \right] dt. \quad \dots (3.14)$$



This is non-negative if and only if  $x_i \geq x_j$ . Therefore (3.13) is proved.

Similarly,  $Af_0(\tilde{Z}) \geq f_i(\tilde{Z}|a)$  is equivalent to

$$\begin{aligned} & Ag_0(\tilde{X}) - g_i(\tilde{X}|a) \\ &= C |x_{N_1-1}^{(1)} - x_{N_1}^{(1)}|^{M+1} \int_0^\infty t^M \exp\left(-\frac{t^2}{2}(N_1+N_2+N_3)S^2\right) \times \\ &\quad \times \left[A - \exp\left[-\frac{1}{2}(N_1+N_2)S_m^2(a)\right] \exp\left(at(x_i - \bar{x})\right)\right] dt. \quad \dots (3.15) \end{aligned}$$

By a simple transformation this is equivalent to

$$\begin{aligned} & \int_0^\infty \xi^M \exp\left[-\frac{1}{2}(N_1+N_2+N_3)\xi^2\right] \times \\ & \times \left[A - \exp\left(-\frac{(N_1+N_2+N_3)}{2} S_m^2(a)\right) \exp\left(a\xi \frac{x_i - \bar{x}}{S}\right)\right] d\xi > 0. \quad \dots (3.16) \end{aligned}$$

As the integral in (3.16) is a continuous increasing function of  $A$  and a continuous decreasing function of  $(x_i - \bar{x})/S$ , for any  $a$  and  $p$  there exists a positive number  $A$  such that (3.16) holds if and only if  $(x_i - \bar{x})/S < \lambda_p$ . Therefore (3.12) is proved. This concludes the proof.

#### 4. AN OPTIMUM DECISION PROCEDURE WHERE $\sigma$ IS KNOWN

In case when  $\sigma$  is known to us, everything becomes quite simple. Under the same notation as in § 2, we can easily see that the procedure will not depend on the values of  $x_1^{(3)} - x_{N_3}^{(3)}$  but it will depend only on  $x_1^{(1)}, \dots, x_{N_1}^{(1)}, x_1^{(2)}, \dots, x_{N_2}^{(2)}$ . Naturally we must cancel the condition (2°). Our optimum solution is found to be the following:

$$\begin{aligned} & \text{If } \frac{x_M - \bar{x}}{\sigma} < \lambda_p, \text{ select } D_0, \\ & \text{If } \frac{x_M - \bar{x}}{\sigma} > \lambda_p \text{ select } D_M, \end{aligned} \quad \dots (4.1)$$

where  $\lambda_p$  is a constant which depends on  $N_1, N_2$  and  $p$ .

We are not going into the details of the derivation of the optimum solution, as it is exactly similar to that of § 3.

In this case, our decision procedure will depend only on  $N_1 + N_2 - 1$  values of  $y_1^{(1)}, \dots, y_{N_1-1}^{(1)}$  and  $y_1^{(2)}, \dots, y_{N_2}^{(2)}$  defined by (3.1). Instead of the joint density function  $(\tilde{Z}|M_1, \dots, M_{N_1-1}, m_{N_1})$  which was required in § 3 and given in (3.7), we used the density function of  $y_1^{(1)}, \dots, y_{N_1}^{(1)}, y_1^{(2)}, \dots, y_{N_2}^{(2)}$ .

# ON THE TESTING OF OUTLYING OBSERVATIONS

Let us write it as  $f(\tilde{Y} | M_1, \dots, M_{N_1-1}, m_{N_1})$ . This is obviously given by

$$f(\tilde{Y} | M_1, \dots, M_{N_1-1}, m_{N_1}) \\ = C \exp \left[ -\frac{1}{2} \frac{1}{(N_1 + N_2)} \left\{ (N_1 + N_2) \left( \sum_{\alpha=1}^{N_1-1} (y_{\alpha}^{(1)} - M_{\alpha})^2 + \sum_{\alpha=1}^{N_2} (y_{\alpha}^{(2)} + m_{N_1})^2 \right) - \right. \right. \\ \left. \left. - \left( \sum_{\alpha=1}^{N_1-1} (y_{\alpha}^{(1)} - M_{\alpha}) + \sum_{\alpha=1}^{N_2} (y_{\alpha}^{(2)} + m_{N_1}) \right)^2 \right\} \right] \quad \dots \quad (4.2)$$

where  $C$  is some constant. This is equal to the following function of  $x_1^{(1)}, \dots, x_{N_1}^{(1)}, x_1^{(2)}, \dots, x_{N_2}^{(2)}$ .

$$g(x_1^{(1)}, \dots, x_{N_1}^{(1)}, x_1^{(2)}, \dots, x_{N_2}^{(2)} | m_1, \dots, m_{N_1}) \\ = C \exp \left[ -\frac{1}{2} (N_1 + N_2) \left\{ S_{12}^2 - 2 \left( \sum_{\alpha=1}^{N_1} m_{\alpha} x_{\alpha}^{(1)} - \left( \sum_{\alpha=1}^{N_1} m_{\alpha} \right) \bar{x} \right) + S_m^2 \right\} \right] \quad \dots \quad (4.3)$$

Proceeding as in § 3 we can easily prove that the decision procedure given in (4.1) is optimum. The details are left to the reader.

In the case  $N_2 = 0$  tables of the values of  $\lambda_p$  have been given by Grubbs (1950) and Nair (1948).

## 5. FURTHER DISCUSSIONS

The arguments given in this paper can be generalized in various ways. For instance, if we know that the population mean values of first  $N_1$  observations are all equal to that of the second group of observations or only one is different i.e. shifted to the right or the left, then the optimum solution (in the same sense as before) will be found to be based on the statistic  $|x_M - \bar{x}|/s$  which is the maximum of  $|x_i - \bar{x}|/s$  ( $i = 1, 2, \dots, N_1$ ). On the other hand, if we know that 2 mean values are shifted to the right by the same amount, the optimum procedure will be based on the statistic  $(x_{M_1} + x_{M_2} - 2\bar{x})/s$  which is the maximum of  $(x_i + x_j - 2\bar{x})/s$  ( $i \neq j, i, j = 1, 2, \dots, N_1$ ). This does not seem to coincide with the statistic proposed by Grubbs for testing two outlying observations.

As we should not select our probability set-up and the decision procedure after getting our samples, we must be very careful in using these procedures. Further investigations on the following problem are obviously necessary. Suppose we have a group of observations and we know *a priori* that they are from the same population or from

two different populations and we have only a vague knowledge about the actual values of the parameters of these two possible cases. For instance, in the notations of § 2, we may have the alternation hypotheses that the  $i_1, i_2, \dots, i_{n-1}$  and  $i_n$ -th ( $1 \leq i_1 < i_2 < \dots < i_n \leq N_1$ ,  $n = 1, 2, \dots, N_1$ ) observations are from another normal population with a different mean and the same variance. In this case we have  $2^{N_1} - 1$  alternative hypotheses. Our problem is how to make decision concerning these hypotheses.

## 6. ACKNOWLEDGEMENT

The author is deeply grateful to the Indian Statistical Institute, where this work was completed and to Mr. D. Basu for his kind suggestions.

## REFERENCES

- GRUBBS, F. E. (1950) : . Sample criteria for testing outlying observations. *Ann. Math. Stat.*, **21**, 27-58.
- NAIR, K. R. (1948) : The distribution of the extreme deviate from the sample mean and its studentized form. *Biometrika*, **35**, 118-144.
- PEARSON, E. S. and CHANDRASEKAR, C. (1936) : The efficiency of statistical tools and a criterion for the rejection of outlying observations. *Biometrika*, **28**, 308-320.
- SMIRNOFF, N. V. (1940) : On the estimation of the maximum term in a series of observations. *C. R. (Doklady). Acad. Sci., New Series*, **33**, 346-350.
- THOMPSON, W. R. (1942) : On a criterion of the difference between the extreme observation and the sample mean in samples of  $n$  from a normal universe. *Biometrika*, **32**, 301-310.

*Paper received : September, 1954.*



# ON SOME QUICK DECISION METHODS IN MULTIVARIATE AND UNIVARIATE ANALYSIS

By J. ROY

*Indian Statistical Institute, Calcutta*

## 0. INTRODUCTION

0.1. A statistical procedure optimum in some classical sense amongst all procedures based on the same number of observations may not be an economic procedure when the cost of observation or of computations are taken into account. For instance, the optimum procedure (in some sense) for testing the equality of means of a number of normal populations with the same variance is the method of analysis of variance, but since the computational labour involved is heavy, quality control engineers prefer the method of control charts. If the cost function for various computational methods can be properly formulated, it may be possible to incorporate this in the general decision theory and optimum rules can be obtained. The difficulties involved in such procedures can be easily imagined. In such circumstances, methods that appear to be cheap and easy to apply have their use. The control chart method, for instance, has not yet been proved to be the most economic procedure, but its justification is that it works quite well.

0.2. Methods based on counting rather than measurements have sometimes been used in industrial problems. Stevens (1946) considered the problem of setting up of control charts by "gauging", that is by counting the number of items in a sample with quality characteristic above or falling short of specified limits. To study the errors of a gun, it may be more convenient to count for a round fired the number of shots hitting the different concentric rings with the bull's-eye as centre rather than to measure the co-ordinates of each shot with respect to orthogonal axes with the bull's-eye as origin. In this case, two measurable characteristics are involved. Methods based on counting may therefore be very useful in certain situations, even though they may not be the optimum in the classical sense.

0.3. In this paper to examine some hypothesis about the underlying distribution of some measurable characteristic, we develop methods that are based mostly on counting rather than on measurement. Counting sometimes is cheaper than measurement. Another property of the method is that no new distribution problem has to be solved.

## 1. A TEST OF A SIMPLE HYPOTHESIS AGAINST A SIMPLE ALTERNATIVE BASED ON THE BINOMIAL DISTRIBUTION

1.1. Let  $x$  be a  $p$ -dimensional chance variable with a continuous probability density function  $f$ . The problem is to test the simple hypothesis  $H_0$  that  $f = f_0$  against the simple alternative  $H_1$  that  $f = f_1$  on the basis of a random sample  $x_1, x_2, \dots, x_n$

of the sign  $n$ . The most powerful test when the first kind of error is fixed at  $\alpha$  is given by:

$$\text{reject } H_0 \text{ if } \prod_{i=1}^n f_0(x_i) < \lambda \prod_{i=1}^n f_1(x_i) \quad \dots (1.1)$$

accept  $H_0$  otherwise

where  $\lambda$  is a constant to be so chosen that the first kind of error is  $\alpha$ . The crux of the problem therefore is to derive the sampling distribution of the likelihood—ratio statistic

$$T \equiv \prod_{i=1}^n \{f_0(x_i)/f_1(x_i)\} \quad \dots (1.2)$$

when the hypothesis  $H_0$  is true, from which  $\lambda$  may be determined to ensure

$$\text{Prob } (T < \lambda | H_0) = \alpha. \quad \dots (1.3)$$

1.2. In many situations, however, the sampling distribution of  $T$  may be very complicated and the evaluation of the percentage point still more complicated, or the computation of the statistic  $T$  itself may be too difficult, or measurement of the  $p$  variables may be inconvenient or costly. Under such circumstances, one may not like to use the test based on the statistic  $T$  even though it is the most powerful one.

The alternative method that we suggest here, though less powerful than the classical method, has certain advantages. The method is based on counting and no new distribution problem has to be solved.

1.3. Let  $\omega$  be a sub-set of the  $p$ -dimensional Euclidean space such that

$$\pi_1 > \pi_0 > 0 \quad \dots (1.4)$$

where

$$\pi_i = \text{Prob. } (x \in \omega | H_i) \quad \dots (1.5)$$

$$i = 0, 1.$$

Let a pseudo-variate  $y_i$  be defined this way

$$\begin{aligned} y_i &= 1 \quad \text{if } x_i \in \omega \\ &= 0 \quad \text{otherwise} \end{aligned} \quad \dots (1.6)$$

Let

$$d = \sum_{i=1}^n y_i. \quad \dots (1.7)$$

Then

$$\text{Prob. } (d = x | H_i) = \binom{n}{x} \pi_i^x (1 - \pi_i)^{n-x} \quad i = 0, 1. \quad \dots (1.8)$$

The statistic  $d$  can therefore be used to test the hypothesis  $H_0$  against the alternative  $H_1$ .

# QUICK DECISION METHODS IN MULTIVARIATE AND UNIVARIATE ANALYSIS

Let  $c$  be the smallest integer to satisfy

$$\sum_{x=c+1}^n \binom{n}{x} \pi_0^x (1-\pi_0)^{n-x} \leq \alpha. \quad \dots (1.9)$$

Then the test procedure is:

$$\text{reject } H_0 \text{ if: } d > c \quad \dots (1.10)$$

otherwise accept  $H_0$ .

The power of this test is given by

$$\begin{aligned} \text{Prob}(d > c | H_1) \\ &= \sum_{x=c+1}^n \binom{n}{x} \pi_1^x (1-\pi_1)^{n-x} \quad \dots (1.11) \\ &= \beta \text{ (say).} \end{aligned}$$

It should be noted that  $\beta$  is an increasing function of  $\pi_1$  for

$$\frac{d\beta}{d\pi_1} = \frac{m - n\pi_1\beta}{\pi_1(1-\pi_1)}$$

where

$$\begin{aligned} m &= \sum_{x=c+1}^n x \binom{n}{x} \pi_1^x (1-\pi_1)^{n-x} > n\pi_1\beta \\ \frac{d\beta}{d\pi_1} &> 0, \quad \dots (1.12) \end{aligned}$$

and therefore

The test is therefore unbiased, and uniformly so far all simple alternatives  $H$ : for which

$$\text{Prob.}(x \in \omega | H) > \pi_0. \quad \dots (1.13)$$

## 2. THE MOST POWERFUL BINOMIAL TEST OF A SIMPLE HYPOTHESIS AGAINST A SIMPLE ALTERNATIVE

2.1 The power of the binomial test discussed in § 1 depends on  $\omega$  and the question that naturally arises is: How should  $\omega$  be chosen so that the power is maximised? Below we give a partial solution to this problem which states that the optimum  $\omega$  must be a member of a particular class.

2.2 Theorem: Let  $\omega$  be a given sub-set of the  $p$ -dimensional Euclidean space satisfying (1.4) and (1.5). Then under certain simple condition it is possible to find a sub-set  $\omega_0$  belonging to the class:

$$f_0(x) < k f_1(x) \quad \dots (2.1)$$

inside  $\omega_0$ :

such that for the same sample size the binomial test based on  $\omega_0$  is at least as powerful as that based on  $\omega$ .



*Proof:* Choose  $k$  to satisfy:

$$\int_{\omega_0} f_0(x)dx = \pi_0 \quad \dots (2.2)$$

Then from Neyman and Pearson's (1933) fundamental lemma it follows that

$$\pi_1^0 = \int_{\omega_0} f_1(x)dx \geq \int_{\omega} f_1(x)dx = \pi_1. \quad \dots (2.3)$$

But the power of the binomial test is an increasing function of  $\pi_1$ . Therefore the test based on  $\omega_0$  is atleast as powerful as that based on  $\omega$ .

2.3. The problem therefore reduces to that of finding an optimum values for  $k$ . If one wishes to maximise the difference between  $\pi_1$  and  $\pi_0$  the solution is  $k = 1$ , but this by itself will not ensure maximum power. It has not been possible to get a general solution for a fixed sample size. However, if  $n$  is so large that the normal approximation to the binomial is satisfactory, the power of the test is approximately given by

$$\beta = \phi \left\{ \tau_\alpha \sqrt{\frac{\pi_0(1-\pi_0)}{\pi_1(1-\pi_1)}} - \sqrt{n} \frac{\pi_1 - \pi_0}{\sqrt{\pi_1(1-\pi_1)}} \right\} \quad \dots (2.4)$$

where

$$\phi(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \quad \dots (2.5)$$

and  $\tau_\alpha$  is defined by

$$\phi(\tau_\alpha) = \alpha \quad \dots (2.6)$$

for  $0 < \alpha < 1$ .

Since  $\beta$  increases as the argument of  $\phi$  decreases, for large values of  $n$  the problem is solved if  $k$  is chosen to maximise

$$\frac{\pi_1 - \pi_0}{\sqrt{\pi_1(1-\pi_1)}}.$$

No general solution could be obtained. To maximise the numerator we may take  $k = 1$ . For any given value of  $n$  however, the optimum value of  $k$  may be determined numerically.

2.4. *Example 1:* To test the hypothesis that the mean of a normal population with a known standard deviation  $\sigma$  is  $\mu_0$  against the alternative hypothesis that it is  $\mu_1$ .

# QUICK DECISION METHODS IN MULTIVARIATE AND UNIVARIATE ANALYSIS

Here,

$$\text{inside } \omega_0 : \frac{\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu_0)^2}}{\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu_1)^2}} < k$$

$$\text{or } x > \frac{\mu_0 + \mu_1}{2} + l\sigma \quad \text{if } \mu_1 > \mu_0 \quad \dots (2.7)$$

where  $l$  is a constant to be suitably determined.

Consequently

$$\pi_0 = \phi(l + \frac{1}{2}\delta) \quad \dots (2.8)$$

$$\text{and } \pi_1 = \phi(l - \frac{1}{2}\delta) \quad \dots (2.9)$$

$$\text{where } \delta = \frac{\mu_1 - \mu_0}{\sigma} \quad \dots (2.10)$$

and  $\phi(x)$  is as defined in (2.5).

As a numerical illustration we tabulate below the power  $\beta$  of the test for the case  $n=100$  and  $\delta=0.1$  with  $\alpha=0.05$  for different values of  $l$ . These were computed by using the approximate formula (2.4). The power of the optimum classical test is also presented.

TABLE 2.1 POWER OF DIFFERENT BINOMIAL TESTS AND OF THE CLASSICAL MOST POWERFUL TEST OF THE MEAN OF A NORMAL POPULATION WITH KNOWN STANDARD DEVIATION

$n = 100 \quad \alpha = 0.05 \quad \delta = 0.1$	
test	power (normal approximation)
binomial test with $l =$	
$-0.2$	0.19
$-0.1$	0.20
$0$	0.20
$+0.1$	0.20
$+0.2$	0.19
most powerful classical test	0.26

2.5. *Example 2:* To test the hypothesis that the standard deviation of a normal population with a known mean  $\mu$  is  $\sigma_0$  against the alternative that it is  $\sigma_1$ .

Here

inside  $\omega_0$  :

$$\frac{\frac{1}{\sigma_0\sqrt{2\pi}} e^{-\frac{1}{2\sigma_0^2}(x-\mu)^2}}{\frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{1}{2\sigma_1^2}(x-\mu)^2}} < k$$

or  $(x-\mu)^2 > l^2\sigma_0^2$  if  $\sigma_1^2 > \sigma_0^2$  ... (2.11)

where  $l$  has to be suitably determined.

Then

$$\pi_0 = 2\phi(l) \quad \dots (2.12)$$

$$\pi_1 = 2\phi(l\rho). \quad \dots (2.13)$$

where

$$\rho = \frac{\sigma_0}{\sigma_1}.$$

The following table of values of  $\beta$  for  $n = 100$ ,  $\alpha = 0.05$  and  $\rho = 0.9$  was computed for different values of  $l$ . The power of the classical most powerful test is also presented.

TABLE 2.2. POWER OF DIFFERENT BINOMIAL TESTS AND OF THE CLASSICAL MOST POWERFUL TEST FOR THE STANDARD DEVIATION OF A NORMAL POPULATION WITH A KNOWN MEAN

$n = 100$	$\alpha = 0.05$	$\rho = 0.9$
test	power (normal approximation)	
binomial test with $l =$		
0.2		0.10
0.3		0.13
0.4		0.15
0.5		0.18
0.6		0.17
most powerful classical test		0.47

2.6. *Example 3:* To test the hypothesis that the vector of mean values of a  $p$ -variate normal distribution with known dispersion matrix  $\Sigma$  is  $\mu_0$  against the alternation that it is  $\mu_1$

Here,

inside  $\omega_0$  :

$$\frac{\frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_0)\Sigma^{-1}(x-\mu_0)'}}{\frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_1)\Sigma^{-1}(x-\mu_1)'}} < k$$

or

$$(\mu_1 - \mu_0)\Sigma^{-1}x' > l\Delta + \frac{1}{2}(\mu_1 - \mu_0)\Sigma^{-1}(\mu_1 + \mu_0)' \quad \dots (2.14)$$



# QUICK DECISION METHODS IN MULTIVARIATE AND UNIVARIATE ANALYSIS

$$\Delta^2 = (\mu_1 - \mu_0)\Sigma^{-1}(\mu_1 - \mu_0)' \quad \dots \quad (2.15)$$

where

and  $l$  is a constant to be suitably determined.

Therefore

$$\pi_0 = \phi(l + \frac{1}{2}\Delta) \quad \dots \quad (2.16)$$

$$\pi_1 = \phi(l - \frac{1}{2}\Delta). \quad \dots \quad (2.17)$$

and

The power of this test depends only on  $l$  and Mahalanobis's distance  $\Delta$ . If we consider the case  $n = 100$ ,  $\Delta = 0.1$ ,  $\alpha = 0.05$ , we have already tabulated the power of the binomial test for different values of  $l$  in Table 2.1.

## 3. MINIMISATION OF SIZE OF SAMPLE WHEN BOTH KINDS OF ERROR ARE PRE-ASSIGNED

3.1. Another problem in testing a simple hypothesis against a simple alternative is to determine the smallest sample size to ensure that the power of the test is a pre-assigned quantity  $\beta$ . For any given  $\omega$  we can find the smallest integer  $n$  such that the power of the binomial test based on a sample of size  $n$  is at least  $\beta$ . Of course,  $n$  will depend on the pre-assigned values of  $\alpha$  and  $\beta$  and the region  $\omega$ . The problem is to so choose  $\omega$  that  $n$  is minimised. Here again we need restrict ourselves to regions of the type:

$$f_0 < \lambda f_1$$

and try to determine  $\lambda$  to minimise  $n$  for fixed values of  $\alpha$  and  $\beta$ .

3.2. Using the normal approximation to the binomial distribution, we get the following requirements on  $n$  and  $c$  to ensure that the first kind of error is  $\alpha$  and that the power is  $\beta$ :

$$\frac{c - n\pi_0}{\sqrt{n\pi_0(1-\pi_0)}} = \tau_\alpha$$

$$\frac{c - n\pi_1}{\sqrt{n\pi_1(1-\pi_1)}} = \tau_\beta$$

with  $\tau_\alpha$  and  $\tau_\beta$  defined by (2.6). From this we get

$$\sqrt{n} = \{\tau_\alpha \sqrt{\pi_0(1-\pi_0)} - \tau_\beta \sqrt{\pi_1(1-\pi_1)}\} / (\pi_1 - \pi_0). \quad \dots \quad (3.1)$$

So  $\omega$  has to be so chosen that this quantity is minimised.

3.3. *Example 1:* Suppose it is required to test that the mean of a normal population with a known standard deviation  $\sigma$  is  $\mu_0$  against the alternative that it is  $\mu_1$  at level of significance  $\alpha$  and power  $\beta$ . Then if  $\mu_1 > \mu_0$  and we take

$$x > \frac{\mu_0 + \mu_1}{2} + l\sigma \quad \dots \quad (3.2)$$

inside  $\omega_0$  :

we have

$$\pi_0 = \phi(l + \frac{1}{2}\delta),$$

$$\pi_1 = \phi(l - \frac{1}{2}\delta)$$

where

$$\delta = \frac{\mu_1 - \mu_0}{\sigma}.$$

The problem is to so choose  $l$  as to minimise  $n$  given in (3.1). This was done numerically for  $\alpha = 0.05$  and  $\beta = 0.90$  and  $0.95$  for  $\delta = 0.25, 0.30, 0.35$  and  $0.40$ . The values of  $l$  and  $n$  for the optimum binomial test and the sample size  $n_0$  required for the optimum non-sequential test are presented in the table below:

TABLE 3.1. SAMPLE SIZE REQUIRED TO TEST THE MEAN OF A NORMAL POPULATION WITH KNOWN STANDARD DEVIATION AT 5% LEVEL OF SIGNIFICANCE

alternative $\delta$	$\beta = 0.95$			$\beta = 0.90$		
	binomial test		classical test	binomial test		classical test
	$l$	$n$	$n_0$	$l$	$n$	$n_0$
0.25	0.275	271	174	0.325	214	138
0.30	0.300	188	121	0.350	148	96
0.35	0.375	137	89	0.375	109	70
0.40	0.400	105	68	0.450	83	54

It will be seen that the binomial test requires a sample size about 1.5 times that required for the classical non-sequential test. Consequently if the ratio of the cost per item sampled for the classical test to that for the binomial test is greater than 1.5 the method suggested here should prove more economic.

3.4. *Example 2:* Consider the problem of testing the vector of mean values of a  $p$  variate normal population with a known dispersion matrix. It immediately follows from 2.6 that the sample size required for the test at level of significance  $\alpha$  to attain a power  $\beta$  for an alternative value of the vector of mean values which is at a distance  $\Delta$  (in Mahalanobis's sense) can be read off directly from Table 3.1 (with  $\delta$  replaced by  $\Delta$ ) independently of the number of variates  $p$  involved so long as  $\Delta$  is kept fixed.

#### 4. TEST OF A SIMPLE HYPOTHESIS ABOUT A SINGLE PARAMETER AGAINST A CLASS OF SIMPLE ALTERNATIVES

4.1. The tests considered in the previous sections do not necessarily possess optimum power properties against a sufficiently wide class of simple alternatives. But it is rather straight forward to build up such tests using general methods due to Neyman and Pearson whenever applicable. For instance, if it is required to build up a test which is uniformly unbiased, we may proceed as follows: Suppose that the problem is to test the hypothesis  $H_0$  that the value of the parameter  $\theta$  involved in the probability density function of the chance variable  $x$  (not necessarily unidimensional) is  $\theta_0$ . Then if we can find a region  $\omega$  such that

$$\pi_{\theta_0} = \int_{\omega} f(x, \theta_0) dx < \int_{\omega} f(x, \theta) dx = \pi_{\theta}$$

for all  $\theta \neq \theta_0$ .

## QUICK DECISION METHODS IN MULTIVARIATE AND UNIVARIATE ANALYSIS

it is immediately seen that the binomial test based on  $\omega$  must necessarily be uniformly unbiased. If then we want to build up a binomial test that is uniformly unbiased and most powerful for a particular alternative  $H_1$  that specifies the value  $\theta_1$  for  $\theta$ , by using a line of argument similar to that used in § 2 and Neyman and Pearson's (1933) fundamental lemma it is easy to show that under the usual regularity condition of differentiability within the integral sign, we need restrict ourselves only to regions of the type:

$$\text{inside } \omega : f(x, \theta_1) > \lambda_1 f(x, \theta_0) + \lambda_2 \left\{ \frac{d}{d\theta} f(x, \theta) \right\}_{\theta_0} \quad \dots (4.1)$$

where  $\lambda_1$  and  $\lambda_2$  are undetermined except for the stipulation that

$$\left[ \frac{d}{d\theta} \int_{\omega} f(x, \theta) dx \right]_{\theta_0} = 0 \quad \dots (4.2)$$

This in general ensures  $\pi_{\theta_0} < \pi_{\theta}$  only for values of  $\theta$  in the neighbourhood of  $\theta_0$ , but if it happens to ensure this uniformly for all  $\theta$ , then only we get uniformly unbiased tests.

This restriction (4.2) gives one of the two constants, the other one has to be determined numerically as in the previous section to maximise the power of the binomial test when the value of the parameter is  $\theta_1$  for a fixed sample size  $n$ , or for very large values of  $n$ .

**4.2. Example 1:** Suppose the problem is to examine the hypothesis that the mean of a normal population with known standard deviation  $\sigma$  is  $\mu_0$  against the alternative hypothesis that it is  $\mu_1$  with the stipulation that the search for the most powerful binomial test must be restricted amongst those that are uniformly unbiased. From (4.1) we get after some simplification:

$$\text{inside } \omega : e^{\delta(\tau - \frac{1}{2}\delta)} > \lambda_1 + \lambda_2 \tau \quad \dots (4.3)$$

$$\text{where } \tau = \frac{x - \mu_0}{\sigma},$$

$$\delta = \frac{\mu_1 - \mu_0}{\sigma}$$

and  $\lambda_1$  and  $\lambda_2$  are undetermined for the present. From the convexity property of the exponential function it follows that (4.3) may be written in the form

$$k_1 < \tau < k_2.$$

Outside  $\omega$ :

The restriction (4.2) implies that  $-k_1 = k_2$  and consequently (4.4) may be written in the final form:

$$|\tau| > k$$

inside  $\omega$ :

where  $k$  is a constant to be suitably determined.



It is now easy to see that

$$\pi_0 = \text{Prob.}(x \in \omega | \mu_0) = 2\phi(k)$$

and

$$\pi_1 = \text{Prob.}(x \in \omega | \mu_1) = \phi(k - \delta) + \phi(k + \delta).$$

This incidentally brings out the symmetry of the power function of the binomial test.

For the special case  $n = 100$ ,  $\alpha = .05$ ,  $\delta = 0.4$ , the following table gives the power of the binomial test for different values of  $k$  as also that of the most powerful uniformly unbiased classical test.

TABLE 4.1. POWER OF DIFFERENT UNIFORMLY UNBIASED BINOMIAL TESTS AND OF THE CLASSICAL TEST FOR THE MEAN OF A NORMAL POPULATION WITH KNOWN STANDARD DEVIATION

(n = 100, $\alpha = 0.05$ , $\delta = 0.4$ )	
test	power (normal approximation)
binomial test with $k$	
0.5	0.14
1.0	0.21
1.5	0.25
2.0	0.25
uniformly unbiased classical test	0.98

## 5. ASYMPTOTICALLY LOCALLY MOST POWERFUL ONE SIDED TEST

5.1. An alternative approach in problems of testing a simple hypothesis about a single parameter is to try to maximise the rate of increase of the power function of a test in a neighbourhood (one sided) of the value of the parameter specified by the null hypothesis. A test for which this property holds may be called an one sided locally most powerful test. (Rao & Pati).

5.2. In this section we consider the problem of finding the locally most powerful binomial test when the sample size is large. In the illustrative example consideration is limited only to a very special class of the binomial tests, as a general solution to the problem could not be derived.

5.3. Suppose that  $f$  involves a single parameter and the hypothesis  $H_0$  to be tested is that the value of the parameter is  $\theta_0$ . Then the power  $\beta$  of the binomial test based on the region  $\omega$  when the value of the parameter is  $\theta_1$  is given approximately, for large values of  $n$  by

$$\beta = \phi(z) \quad \dots (5.1)$$

where

$$z = \tau_\alpha \sqrt{\frac{\pi_0(1-\pi_0)}{\pi_1(1-\pi_1)}} - \sqrt{n} \left( \frac{\pi_1 - \pi_0}{\sqrt{\pi_1(1-\pi_1)}} \right) \quad \dots (5.2)$$

where

$$\pi_0 = \text{Prob.}(x \in \omega | \theta_0)$$

$$\pi_1 = \text{Prob.}(x \in \omega | \theta_1).$$



# QUICK DECISION METHODS IN MULTIVARIATE AND UNIVARIATE ANALYSIS

Therefore, the rate of increase of the power at the point  $\theta_1$  is

$$\begin{aligned} \frac{\partial \beta}{\partial \theta_1} &= \frac{\partial \beta}{\partial z} \cdot \frac{\partial z}{\partial \pi_1} \cdot \frac{\partial \pi_1}{\partial \theta_1} \\ &= \left( -\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \right) \left\{ \tau_\alpha \frac{(\pi_1 - \frac{1}{2})\sqrt{\pi_0(1-\pi_0)}}{\pi_1(1-\pi_1)\sqrt{\pi_1(1-\pi_1)}} - \right. \\ &\quad \left. - \sqrt{n} \left( \frac{(\pi_1 - \pi_0)(\pi_1 - \frac{1}{2})}{\pi_1(1-\pi_1)\sqrt{\pi_1(1-\pi_1)}} + \frac{1-\pi_0}{\sqrt{\pi_1(1-\pi_1)}} \right) \right\} \left( \frac{\partial \pi_1}{\partial \theta_1} \right). \end{aligned}$$

Consequently

$$\left( \frac{\partial \beta}{\partial \theta_1} \right)_{\theta_1=\theta_0} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\tau_\alpha^2} \left\{ \sqrt{n} \sqrt{\frac{1-\pi_0}{\pi_0}} - \tau_\alpha \frac{\pi_0 - \frac{1}{2}}{\pi_0(1-\pi_0)} \right\} \left( \frac{\partial \pi_1}{\partial \theta_1} \right)_{\theta_1=\theta_0}. \quad \dots (5.3)$$

The dominant term in this expression is

$$\sqrt{n} \cdot \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\tau_\alpha^2} \sqrt{\frac{1-\pi_0}{\pi_0}} \times \left( \frac{d\pi_1}{d\theta_1} \right)_{\theta_1=\theta_0}. \quad \dots (5.4)$$

The problem therefore is to so choose  $\omega$  that

$$\sqrt{\frac{1-\pi_0}{\pi_0}} \left( \frac{\partial \pi_1}{\partial \theta_1} \right)_{\theta_1=\theta_0} \quad \dots (5.5)$$

is maximised.

But a general solution to this problem could not be derived. However, if  $\theta$  is a location parameter, that is if  $f(x, \theta) \equiv f(x - \theta)$  and the range does not involve  $\theta$  and consideration is limited only to regions of the type

$$x > \theta_0 + k \quad \dots (5.6)$$

$\omega$  :

it is easy to see that

$$\left( \frac{\partial \pi_1}{\partial \theta_1} \right)_{\theta_1=\theta_0} = f(k).$$

Consequently  $k$  has to be chosen to maximise

$$U = \sqrt{\frac{1-\pi_0}{\pi_0}} f(k) \quad \dots (5.7)$$

$$\pi_0 = \int_k^\infty f(t) dt. \quad \dots (5.8)$$

and, here

The condition

$$\frac{dU}{dk} = 0 \quad \text{gives}$$

$$2\pi_0(1-\pi_0)f'(k) + \{f(k)\}^2 = 0. \quad \dots (5.9)$$

It follows that for the problem of testing the mean of a normal population with unit standard deviation the value of  $k$  is a root of

$$2k = \frac{z}{\phi(1-\phi)} \quad \dots (5.10)$$

where

$$z = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}k^2}$$

and

$$\phi = \int_k^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} dt.$$

#### REFERENCES

- NEYMAN, J. and PEARSON E. S. (1933): On the problem of the most efficient tests of statistical hypotheses  
*Phil. Trans. Roy. Soc. A*, 231, 281.
- RAO, C. R. and POTI, S. J. (1946): On locally most powerful tests when alternatives are one sided,  
*Sankhyā*, 7, 439.
- STEVENS, J. (1948): Quality Control by Gauging. *J. Roy. Stat. Soc. Ser. B*.

*Paper received : January, 1955.*

# ON THE METHOD OF OVERLAPPING MAPS IN SAMPLE SURVEYS

By DES RAJ

*Indian Statistical Institute, Calcutta*

## 1. INTRODUCTION

In multipurpose surveys involving the estimation of several characters, it is usually found desirable to select the units with one set of probabilities for estimating one group of characters and with a different set of probabilities for estimating another group of characters. For example, in the National Sample Survey (NSS) of India, population is made the basis for selection for the household enquiry and area for the land utilisation survey. An important problem arising in such a situation is that of designing a suitable selection procedure, so that the sample units (villages, in case of the NSS) for the two types are almost identical or near to one another. Such a procedure will greatly reduce the cost of operations in the field.

Lahiri (1954) has given two methods of selection for the purpose, called the 'serpentine' method and the 'two dimensional' method. He has not, however, entered into the mathematics of the problem. The object of this paper is to present the problem mathematically and offer general solutions.

## 2. FORMULATION OF THE PROBLEM

Suppose a tract contains  $n$  villages. We are required to select a pair of sample villages, one with probabilities

$$\frac{a_1}{G}, \frac{a_2}{G}, \dots, \frac{a_n}{G}$$

proportional to area and the other with probabilities

$$\frac{b_1}{G}, \frac{b_2}{G}, \dots, \frac{b_n}{G}$$

proportional to population in such a manner that the two villages are close to one another, if not identical. Let  $c_{ij}$  be the distance (in some sense) between the  $i$ -th area village and the  $j$ -th population village.

TABLE 1. AMOUNTS OF PROBABILITY MASS TO BE DISTRIBUTED

village no.	by population						total
	1	2	...	$j$	...	$n$	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	$x_{11}$	$x_{12}$	...	$x_{1j}$	...	$x_{1n}$	$a_1$
2	$x_{21}$	$x_{22}$	...	$x_{2j}$	...	$x_{2n}$	$a_2$
...	...	...	...	...	...	...	...
$i$	$x_{i1}$	$x_{i2}$	...	$x_{ij}$	...	$x_{in}$	$a_i$
...	...	...	...	...	...	...	...
$n$	$x_{n1}$	$x_{n2}$	...	$x_{nj}$	...	$x_{nn}$	$a_n$
total	$b_1$	$b_2$	...	$b_j$	...	$b_n$	$G$



Let  $x_{ij}/G$  ( $i, j = 1, \dots, n$ ) be the probability with which the corresponding pair of villages is selected. The problem is to find  $x_{ij}$  such that

$$\sum_{j=1}^n x_{ij} = a_i, \sum_{i=1}^n x_{ij} = b_j; \quad \sum_1^n a_i = \sum_1^n b_j = G; \quad x_{ij} \geq 0$$

and

$$Z = \sum \sum c_{ij} x_{ij} \text{ is minimised.}$$

Stated thus, this is the familiar 'transportation problem' (Koopmans, 1951) in linear programming. The solution by the simplex method due to Dantzig is given in the book referred to above.

Starting with the arbitrary basic solution of Dantzig, one gets the final solution in a finite number of stages. If the villages are arranged in a serpentine fashion and the same arrangement is used for area as well as for population, it is interesting to see that Lahiri's 'serpentine method' is the same as Dantzig's arbitrary basic solution with which the iterative process starts.

### 3. AN ILLUSTRATION

As an illustration of the method, we consider the following ten villages, the map of which is given in Fig. 1 below:

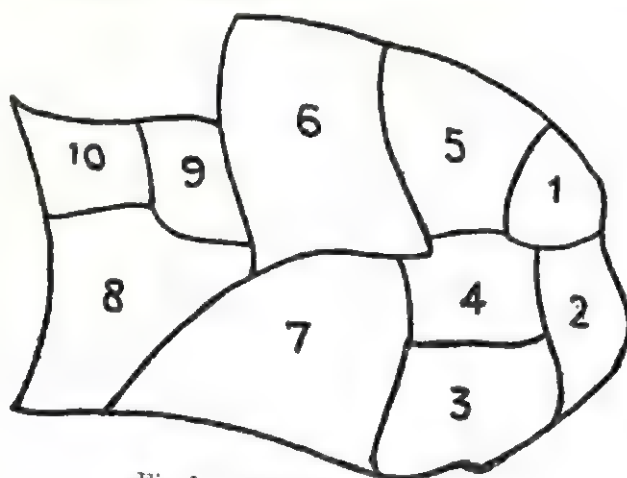


Fig 1. Map of ten villages

The villages are numbered in a serpentine fashion and their areas and populations are given in the table below:

TABLE 2. AREA AND POPULATION

village no.	area	population	village no.	area	population
(1)	(2)	(3)	(1)	(2)	(3)
1	3	8	6	15	20
2	4	5	7	15	10
3	6	10	8	11	5
4	5	5	9	5	10
5	11	6	10	5	1

# ON THE METHOD OF OVERLAPPING MAPS IN SAMPLE SURVEYS

Measuring the actual geographical distances (from the centres) between the villages, the cost matrix is obtained as the following:

TABLE 3. COST MATRIX  $c_{ij}$

village no.	by population										
	1	2	3	4	5	6	7	8	9	10	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	
1	0	5	10	5	5	11	14	20	15	21	
2	5	0	7	5	8	13	13	21	18	23	
3	10	7	0	6	11	14	8	17	16	20	
4	5	5	6	0	6	9	9	16	13	18	
5	5	8	11	6	0	6	11	16	12	17	
by area	6	11	13	14	9	6	0	10	11	5	10
	7	14	13	8	9	11	10	0	9	9	12
	8	20	21	17	16	16	11	9	0	6	5
	9	15	18	16	13	12	5	9	6	0	5
	10	21	23	20	18	17	10	12	5	5	0

The tables below give the arbitrary basic solution (of Dantzig), the final solution and salient features of the iterative process.

TABLE 4. ARBITRARY BASIC SOLUTION

village no.	by population										total
	1	2	3	4	5	6	7	8	9	10	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
											3
1	3										4
2	4										6
3	1	5	0								5
4			5								11
5			5	5	1						
by area					5	10					15
	6					10	5				15
	7						5	5	1		11
	8								5		5
	9								4	1	5
	10										
total	8	5	10	5	6	20	10	5	10	1	80

TABLE 5. FINAL SOLUTION

village no.	by population									
	1	2	3	4	5	6	7	8	9	10
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
1	3									
2		4								
3			6							
4		0		5						
5	5	0			6					
by area										
6						15				
7		1	4			0	10			
8						5		5	1	
9									5	
10									4	1

TABLE 6. BRIEF DETAILS OF THE ITERATIVE PROCESS

step (i)	$Z_i = Z_{i-1} - M_{i-1} \theta_{i-1}$	$M_i = \max.(\bar{c}_{ij} - c_{ij})$	$\theta_i$
(1)	(2)	(3)	(4)
1	381	26	1
2	355	28	4
3	253	22	0
4	243	23	0
5	243	13	1
6	230	10	4
7	190	8	4
8	158	16	0
9	158	13	0
10	158	7	1
11	151	0	

It is found by the author that in problems of this type where the cost matrix contains all zeros in the diagonal, instead of starting with Dantzig's solution, which is very inefficient in this case, one should start with any solution with the maximum mass in the diagonal (obtained by putting  $\min(a_i, b_i)$  in the diagonal). In fact, in this example, seven steps were necessary to get at a solution with the maximum mass in the diagonal and then only four further steps were required to get the optimum solution.

# ON THE METHOD OF OVERLAPPING MAPS IN SAMPLE SURVEYS

It is of interest to note in this connection that Lahiri's two-dimensional solution (in table 7 below), obtained by inspection, is almost as good as the optimum since  $Z = 152$  in this case.

TABLE 7. LAHIRI'S TWO-DIMENSIONAL SOLUTION

village no.	by population									
	1	2	3	4	5	6	7	8	9	10
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
1	3									
2		4								
3			6							
4		1		4						
5	5				5					
by area						15				
6			4	1			10			
7						4		5	2	
8									5	
9						1			3	1
10										

## 4. THE OPTIMUM CHARACTER OF THE SERPENTINE METHOD

In the illustration given above the distance between two villages is defined as the actual geographical distance between the centres of the villages. To simplify the problem we may define 'distance' as the difference between the serial numbers of the villages when the villages have been numbered in a serpentine fashion. We shall now show that, with this definition of distance, Dantzig's arbitrary basic solution (which is the same as Lahiri's serpentine method) gives the optimum solution in the sense that it minimises the expected 'distance'. Without any loss of generality we shall take the case of  $n = 5$  villages with areas and populations given in Table 1 before.

Let Dantzig's arbitrary solution be given by Table 8. The essential feature of this solution is that if one travels from the first to the last cell containing the basis, the path is a continuous one and is always parallel to the sides.

TABLE 8. ARBITRARY SOLUTION

village no.	by population					total
	1	2	3	4	5	
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	$x_{11}$		$x_{23}$			$a_1$
2	$x_{21}$	$x_{22}$	$x_{33}$	$x_{34}$		$a_2$
by area				$x_{44}$		$a_3$
3				$x_{54}$	$x_{55}$	$a_4$
4						$a_5$
5						$G$
total	$b_1$	$b_2$	$b_3$	$b_4$	$b_5$	



TABLE 9. COST MATRIX  $c_{ij}$ 

village no.	by population				
	1	2	3	4	5
(1)	(2)	(3)	(4)	(5)	(6)
1	0	1	2	3	4
2	1	0	1	2	3
by area 3	2	1	0	1	2
4	3	2	1	0	1
5	4	3	2	1	0

With this definition of distance, the (direct) cost matrix is given by Table 9. To obtain the indirect cost matrix  $\bar{c}_{ij}$ , we have  $\bar{c}_{ij} = c_{ij}$  for any element  $(i, j)$  appearing in the basis. The other elements  $\bar{c}_{ij}$  are given by  $\bar{c}_{ij} = u_i + v_j$  where  $u_i$  and  $v_j$  are to be determined from the elements in the basis.

Let  $u_1 = c_{11}$  and  $v_1 = 0$ . Then we have in succession

$$u_2 = c_{21} - v_1 = c_{21},$$

$$v_2 = c_{22} - u_2 = c_{22} - c_{21},$$

$$v_3 = c_{23} - u_2 = c_{23} - c_{21},$$

$$u_3 = c_{33} - u_3 = c_{33} + c_{21} - c_{23},$$

$$v_4 = c_{34} - u_3 = c_{34} - c_{33} + c_{23} - c_{21},$$

$$u_4 = c_{44} - v_4 = c_{44} - c_{34} + c_{33} + c_{21} - c_{23},$$

$$u_5 = c_{54} - v_4 = c_{54} - c_{34} + c_{33} - c_{23} + c_{21},$$

$$v_5 = c_{55} - u_5 = c_{55} - c_{54} + c_{34} - c_{33} + c_{23} - c_{21}.$$

TABLE 10. COST MATRIX  $\bar{c}_{ij}$  CORRESPONDING TO BASIS

village no.	by population				
	1	2	3	4	5
(1)	(2)	(3)	(3)	(5)	(6)
1	$c_{11}$				
2	$c_{21}$	$c_{22}$	$c_{23}$		
by area 3			$c_{33}$	$c_{34}$	
4				$c_{44}$	
5				$c_{54}$	$c_{55}$

Substituting  $c_{ij} = |i-j|$ , we have

$$u_1 = 0, u_2 = 1, u_3 = 0, u_4 = -1, u_5 = 0$$

and

$$v_1 = 0, v_2 = -1, v_3 = 0, v_4 = 1, v_5 = 0.$$

# ON THE METHOD OF OVERLAPPING MAPS IN SAMPTA SURVEYS

Hence the indirect cost matrix  $\bar{c}_{ij}$  is as given in Table 11.

TABLE 11. INDIRECT COST MATRIX  $\bar{c}_{ij}$

village no.	by population				
	1	2	3	4	5
(1)	(2)	(3)	(4)	(5)	(6)
1	0	-1	0	1	0
2	1	0	1	2	1
3	0	-1	0	1	0
by area 4	-1	-2	-1	0	-1
5	0	-1	0	1	0

Comparing this with the direct cost matrix  $c_{ij}$  of Table 9, we see that

$$M = \max (\bar{c}_{ij} - c_{ij}) = 0$$

so that the solution obtained is optimum.

## 5. SOME FURTHER REMARKS ON THE SERPENTINE METHOD

A number of interesting conclusions emerge from the result proved above. We have proved that if the villages are arranged in a serpentine fashion, the serpentine method gives the optimum solution. But it is possible to arrange the villages in a serpentine fashion in a number of ways. Since for any serpentine arrangement the marginal totals  $a_i$ 's and  $b_j$ 's would remain unchanged (they may occur in a different order), the same minimum value of the expected 'distance' is achieved, in whatever manner the villages be arranged provided the arrangement is serpentine. This proves that the method given by Lahiri (1954) namely 'when the serpentine arrangement is more or less at our choice we should endeavour to arrange the villages in such a manner that the cumulative density fluctuates as frequently as possible about the tehsil density', cannot improve the situation. Any serpentine arrangement will do as well.

## 6. THE PROBLEM OF KEYFITZ

Mention must be made here of the work of Keyfitz (1951) who considers the case when first stage units within strata are to be selected with different probabilities at two successive occasions, so that the probability of having identical first stage units at the two occasions is maximised. Stated mathematically, the problem is the same as discussed in § 2 where the cost matrix is given in Table 12 below:

TABLE 12. COST MATRIX IN KEYFITZ' PROBLEM

unit no.	1951				
	1	2	3	...	$n$
(1)	(2)	(3)	(4)	(5)	(6)
1	0	1	1	...	1
2	1	0	1	...	1
3	1	1	0	...	1
$\vdots$					
$n$	1	1	1	...	0

By the general theory (and obviously enough) the optimum solution consists in putting in the diagonals as much mass as possible, viz.,  $\min(a_i, b_i)$  in the  $(i, i)$  cell. We shall now show that the somewhat unwieldy selection procedure given by Keyfitz is equivalent to the simple procedure given by us. To take a concrete case, let  $n = 5$  in Table 1 and  $a_1 > b_1, a_2 > b_2, a_3 > b_3, a_4 < b_4, a_5 < b_5$  (there is no loss of generality involved here). Then the probability of getting identical villages in our method is given by

$$\frac{1}{G} \sum_{i=1}^5 \min(a_i, b_i) = \frac{1}{G} (b_1 + b_2 + b_3 + a_4 + a_5).$$

By Keyfitz' method, probability of getting identical villages is equal to

$$\begin{aligned} 1 - \frac{1}{G} (b_4 + b_5 - a_4 - a_5) \\ = 1 - \frac{1}{G} (G - b_1 - b_2 - b_3 - a_4 - a_5) = \frac{1}{G} (b_1 + b_2 + b_3 + a_4 + a_5) \end{aligned}$$

so that the two methods are equivalent i.e. Keyfitz' method is also optimum. It may be noticed here that Keyfitz is thinking of a situation where the selection is to be made at two different occasions viz.,  $a$ 's are the sizes for 1950 while  $b$ 's are the sizes for 1951. In such a case, we will select a village with probabilities proportional to  $a$ 's in 1950 (say village 1 is selected in the sample).

Then at the second occasion we will select a village with probabilities proportional to  $x_{11}/a_1, x_{12}/a_1, \dots, x_{15}/a_1$  so that the overall probability that a combination (say) (1, 3) is selected is

$$\frac{a_1}{G} \times \frac{x_{13}}{a_1} = \frac{x_{13}}{G} \text{ as desired in the method.}$$

## 7. THE PROBLEM OF GOODMAN AND KISH

Another related problem occurs in stratified sampling. In the usual type of stratified sampling, units within one stratum are selected independently of those within another stratum. But more generally, one may select a pair of units, one from each stratum, in a dependent way such that certain preferred types of pairs have a higher probability of selection and consequently others have a lower probability of selection. Such a problem has been considered by Goodman and Kish (1950). Their population consists of two strata, one containing 3 coastal and 3 inland units while the other contains five units of which one is coastal and the others are inland. Two units have to be selected, one from each stratum, with assigned probabilities within strata. The object is to maximise the probability of selecting one coastal and one inland unit. It is obvious that this problem easily reduces to the one considered before. The cost matrix is given in Table 13 below:

# ON THE METHOD OF OVERLAPPING MAPS IN SAMPLE SURVEYS

TABLE 13. COST MATRIX IN GOODMAN AND KISH'S PROBLEM

		stratum 1					
		coastal			inland		
units		B	C	F	A	D	E
(1)		(2)	(3)	(4)	(5)	(6)	(7)
stratum 2	inland a	0	0	0	1	1	1
	b	0	0	0	1	1	1
	c	0	0	0	1	1	1
	e	0	0	0	1	1	1
	coastal d	1	1	1	0	0	0

The optimum solution, giving the probabilities with which pairs of units be selected, is presented in Table 14.

TABLE 14. OPTIMUM SOLUTION IN GOODMAN AND KISH'S PROBLEM

units	stratum 1						total
	B	C	F	A	D	E	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
stratum 2	a	15			0		15
	b	0	10	20			30
	c				10	0	10
	e					20	25
	d						20
total	15	10	20	10	20	25	100

This type of problem, however, raises new issues. In the previous problem we were concerned with the estimation of several characters and a particular scheme of selection did not affect the variances of the estimates of individual characters so long as the units within strata were selected with certain assigned probabilities. But in this case since the same character is under investigation in the two strata, the variance of the final estimate will be affected due to the dependent manner in which pairs of units are selected. To be more specific, let there be two strata of sizes  $N_1$  and  $N_2$  respectively. From the first stratum a unit is to be selected with probabilities  $p_i (i=1, \dots, N_1)$  and another unit is to be selected from stratum two with probabilities  $p'_j (j=1, 2, \dots, N_2)$  in such a way that the expected cost (of travel) is minimised. Suppose  $p_{ij} (i=1, \dots, N_1; j=1, 2, \dots, N_2)$  is the optimum set of probabilities for selecting the  $i$ -th unit from the first stratum and the  $j$ -th unit from the second stratum.

Then 
$$\hat{y}_{dep} = \frac{y_i}{p_i} + \frac{y'_j}{p'_j}$$

is an unbiased estimate of the population total.

Also 
$$V(\hat{y}_{dep}) = \sum \frac{y_i^2}{p_i} + \sum \frac{y_j'^2}{p_j'} + 2 \sum \sum p_{ij} \frac{y_i}{p_i} \frac{y'_j}{p'_j} - Y^2.$$



In case selection within one stratum is independent of that within another, we have

$$\hat{y}_{ind} = \frac{y_i}{p_i} + \frac{y'_j}{p'_j}.$$

$$V(\hat{y}_{ind}) = \sum \frac{y_i^2}{p_i} + \sum \frac{y_j'^2}{p_j'} - (Y_1^2 + Y_2^2)$$

so that

$$V_{dep} - V_{ind} = 2 \sum \sum \frac{p_{ij} - p_i p'_j}{p_i p'_j} y_i y'_j.$$

One cannot establish that  $V_{dep}$  is always smaller than  $V_{ind}$  so that the possibility is there that the new method, though decreasing the travel cost, may increase the variance of the estimate.

Another point worth mentioning is the estimation of error variance in such designs. Obviously enough, it is not possible to get unbiased estimates of the sampling error even in case of independent sampling within strata by selecting only one unit from each stratum. If, however, a duplicate set of units is selected (in the same way as before), it is possible to obtain unbiased estimates of the sampling error. If  $\hat{y}_{1dep}$  and  $\hat{y}_{2dep}$  denote the population total estimates based on the first and the second set and  $\hat{y}'_{dep}$  denotes the pooled estimate, we have

$$\hat{y}'_{dep} = \frac{1}{2}(\hat{y}_{1dep} + \hat{y}_{2dep}),$$

$$V(\hat{y}'_{dep}) = \frac{1}{2}V(\hat{y}_{dep}),$$

$$\hat{V}(\hat{y}'_{dep}) = \left( \frac{\hat{y}_{1dep} - \hat{y}_{2dep}}{2} \right)^2.$$

In case sampling within one stratum is independent of that within another, we have the comparable quantities:

$$\hat{y}'_{ind} = \frac{1}{2}(\hat{y}_{1ind} + \hat{y}_{2ind}),$$

$$V(\hat{y}'_{ind}) = \frac{1}{2}V(\hat{y}_{ind}),$$

$$\hat{V}(\hat{y}'_{ind}) = \left( \frac{\hat{y}_{1ind} - \hat{y}_{2ind}}{2} \right)^2.$$

#### REFERENCES

1. GOODMAN, R. AND KISH, L. (1950): Controlled selection—a technique in probability sampling. *J. Amer. Stat. Ass.*, **45**, 350-72.
2. KEYFITZ, N. (1951): Sampling with probabilities proportional to size — adjustment for changes in the probabilities. *J. Amer. Stat. Ass.*, **46**, 105-109.
3. KOOPMANS, T. C. (1951): *Activity Analysis of Production and Allocation*, John Wiley and Sons, New York, 359-373.
4. LAHIRI, D. B. (1954): Technical paper on some aspects of the development of the sample design. *The National Sample Survey No. 5, Ministry of Finance, Government of India.*

Paper received ; March, 1955.

## BOOK REVIEWS

**Survey of Fertility and Mortality in Poona District :** By V. M. Dandekar and K. Dandekar; Gokhale Institute of Politics and Economics, Publication No. 27; Price Rs. 5/-.

This is the first publication on the work done in the Section for Demography and Population Studies, established in the Gokhale Institute in 1951 with a donation from the Rockefeller Foundation. In contemporary India study of the dynamics of population growth has not received the attention it deserves and it is heartening to see a new contribution on the subject. The authors have devoted a good deal of labour and thought to analyse and present the results, and to interpret them objectively. The difficulties with which an operator has to grapple in the field are brought into relief in this work. The book is thus a welcome and valuable addition to our limited store of knowledge and experience in the sector. Some points in the design and methods adopted are however open to criticism and one could not agree with the views expressed by the authors at some other places.

The main objectives of the Survey are stated in the foreword (p.ii) by Dr. Gadgil, the Director of the Institute, as collection of information regarding births and deaths and the attitude towards family planning. The Institute was however disappointed in the mortality data collected, scrapped the Survey data and fell back on the municipal birth and death registration and 1951 census population figures to produce life tables for Poona City. Birth and death registration statistics in India are generally known to be very much defective. One would view them with grave suspicion even for obtaining overall death rates; but in this publication, they have been accepted on face value, without any adjustments, for constructing a set of age specific mortality rates.

The primary problem facing a demographer in framing a detailed life table from raw data is the testing of the age returns for the so-called 'integer bias' and devising means to cure the fault. The problem has been rather summarily disposed of in para A.4 and quinary age groups with end digits 3-7 : 8-2 adopted without testing the efficiency of the system, apparently in the belief that since the digits 0 and 5 have the maximum pull, quinary age groups centred round them would be most efficient. It is not possible to suggest the most efficient system of age grouping of the medium as the basic data in individual ages has not been given in the report, but from general experience it looks probable that the 2-6 : 7-1 or even the straight-forward 0-4 : 5-9 system might have done better. With the system of grouping adopted, age concentration shows up as early as in age-groups 8-12 and 18-22. A look at the Tables 2.11 and A.2, with their marked heaping at the age-groups round end-digit 0, should have at once cast doubts in the mind of any operator.

From the registered deaths those occurring in public hospitals to non-residents were rightly excluded, but it should be realised that the other side of the account, of migration

to home outside City on attack of sickness and death there, has also been missed. It is not clear how the population commuting to the City has been treated. As to the under-representation of deaths in the actual Survey data (Chapter VI) it is also relevant to remember that shifts in relation categories and in exposure quantities are inherent in the design in the event of death of the household-head during the previous year. It is advisable to take the average of deaths over 3 years, say, around the census or survey date to get working practical rates of mortality, but only a single year's deaths have been used by the authors. It is thus a matter of guess how much of the resulting improvement in the expectation of life in the Tables A.7 and A.8 arises from under-registration and how much from chance short period fluctuation. But such detailed life tables for an open City like Poona based on a single year's experience is of ambiguous practical value anyway, and no new techniques which could be of theoretical interest have been introduced.

Coming next to the Survey proper and the design of the Survey, it is to be noticed at the outset that the frame for Poona City (and other towns included in the Survey) is the ration register. The inflation in ration lists, both in aggregate numbers and in ages (at the crucial age line dividing children from adults) is notorious. While inflations are more numerous in certain sections, omissions are perhaps likely in others, specially among the agricultural classes and the lower income groups in the fringes of towns. Table 3.1 gives the number of families registered in Poona City frame as 1,40,098 while Table 3.2 gives the number of households in the City according to 1951 census as only 1,01,460. There is thus about 40% inflation in households with the census enumeration as standard. The total number of individuals on ration lists has not been given, but the total population of Poona City at 1951 census is quoted in Table 2.1 as 4,80,982; distributed in the families registered at the ration shops, this works out to less than 3.5 members per family as against over 4.7 members per census household.

The inordinate difficulty experienced in the Survey (para 3.16) in tracing the households selected was only natural in this background. Ward names in Tables 3.1 and 3.2 do not all tally (Guruwar, Nagesh, Shivajinagar, Nihal, Vetal, Ghorpade) and it has not been stated how the selection made in one ward in Table 3.1 has been allocated in another ward in Table 3.2. Attention has not been drawn in the body of the report even to this fundamental inflation and discrepancy in the frame. The under-registration of certain wards in the final sample mentioned in para 3.17 and the serious fault in the frame pointed out above have no doubt affected the results in a number of ways. The sample towns were apparently selected by a haphazard procedure (para 3.18) with bias for rationed towns and so were the sample villages with bias for villages with population less than 1,500 (para 3.20). On the top of it, the selection of the sample households in the villages was left to the choice of the investigators, provided they covered all sections of the village site (para 3.21). Incidentally, it is interesting to find a small rural population in Poona City area, as shown in Table 2.2.

The 'biological family' has been chosen as the unit of investigation. The authors were very forthright in condemning the 'household' as the unit for such surveys. In para 3.6 they even go to the extent of saying that some definitions of the households as unit work only to the extent that they are kept vague and any clarification of the term only tends to defeat its purpose. Though border-line conceptual difficulties do arise, there is a definition (and perhaps there are more than one) of a *messing-cum-resident household unit* which



## BOOK REVIEW

is as satisfactory and precise as the definition of any other unit. That may not be known to the authors and their enthusiasm to run down the household as unit might have originated from this situation. But the authors had themselves to fall back on the basic unit of household to determine their unit of 'biological family' (para 3.8). They have, however, not defined the 'household' used by them in the Survey to reach the ultimate unit. Para 3.8 provides that if there were any adult male in a selected household with de facto female head, he was to be treated as the head and his bio-family was enumerated. It is not known what happened if a servant was the only adult male member of the 'household' (or if a servant could never be a member of the 'household').

The report again does not indicate how step-relations were to be treated. The schedule does not provide for polyandric marriages, though it may be of small significance in the particular geographical area covered. If the head had more than one wife living, a birth during the year only to the latter wife and her fertility history were presumably to be recorded; if the subsequent marriage was occasioned by the barrenness of the first wife, the resulting fertility rates would have been overestimated to that extent. It is also not clear if a birth during the year to a wife dead on the date of investigation was to be counted. Possibly these points were dealt with in the instructions to the field staff, but they are mentioned here as they are important in assessing the proper import of the resulting fertility rates. Subsidiary families, say a first cousin, his sibs, wife and children, or an employee and his relations staying in the same household had no chance of inclusion in the Survey. The characteristics including fertility derived from the Survey results, therefore, relate only to a population of families of which the head is the 'head' of a 'household'. The bio-family unit is overlapping and environmental influences are likely to be diffused and obliterated in working with it.

The family composition statistics clearly brings to surface the bias in the design and the procedure. It has been mentioned in the report that the members of families with bigger number of adult male members (and more of them living separate) had greater chances of selection as sample (para 4.15), but the implications of such bias are not fully explained. There were not only greater chances of selection of mothers of big families in the sample, but chances of repeated selections in different bio-families (if there were separate households of sons). It is evident that a sample so selected will be biased in favour of bigger bio-families and the historical fertility rates would be exaggerated. The actual number of children born to the generation of 'mother's was likely to be affected most by this bias, and the lift in the historical fertility curve in graphs nos. 2 and 4 at durations over 20 might at least be partly explained by this.

From Table 4.17 it would appear that while a thousand wives at marriage duration 1 had only 50 children during the year preceding the Survey, a thousand wives at marriage duration 2 had 240 children during the same year with total children born to date 640. It follows that the 1000 women at marriage duration 2 had 640—240 or 400 children in the 1 year duration of their marriage. The productive performance during the 1 year duration of their marriage of these wives at marriage duration 2, is thus eight times the performance of the wives married in the subsequent year. These results are definitely inconsistent and confer an initial handicap to the series of cumulated duration specific fertility rates. This inconsistency has not at all been examined by the authors. A probable explanation might be that the small families with recently married wives and newly born babies were missed from sample.



The curve of cumulated fertility rates for Poona City (graph 2) again loses ground sharply at marriage durations 6-8. It is not unlikely that this sudden and sharp rise in the historical fertility curve might have partially resulted from the motive to describe the children as adults for rationing purpose. This suspicion is reinforced when it is seen that the break in the historical fertility curve from the cumulated fertility curve starts much later, at about marriage duration 15, for the towns and villages (graph 4). From Table A.4 it moreover appears that the total births registered in the year ended August 1951 at the Poona Corporation were 13,966, which is over 30% in excess of the 10,796 births registered in the year ended August 1950. Perhaps much of the reference period in the present Survey fell in the later year, 1951. There is thus some external evidence that the year from which the cumulated fertility rates are built up was not an unfavourable year for births. Two deaths to sons and 7 deaths to daughters, under 1 year of age, were reported in the Survey as against 15 death, expected in each category on municipal registration basis (para 6.3); this was indicative of serious under-reporting of dead infants.

The figures of total number of children born are from the category of women living, with husbands also living, and association between mortality and fertility may be a contributory cause of the deviation between the historical and cumulated fertility curves. Further analysis therefore appears necessary before the conclusion drawn in para 4.25 that the women were subjected to higher actual fertility at various durations than those derived from births occurring in the previous year could be accepted. A considerable portion of the report has been devoted to the tests of significance of the effects of various factors on fertility rates. Chi-square tests applied failed to bring out any significance of most of the factors chosen. The lumping together of the towns and villages in one set of data and the spread of the bio-family unit across environmental sectors had presumably resulted in some diffusion and mixing up of the differentials. An attempt might also have been made at studies with control and isolation of some factors in evaluating the influence of the important factors.

It is important to note that the Institute has now acquired the experience that local field investigators were not the suitable agency in small towns (and villages) as questioning the acquainted was difficult (para 3.2) and embarrassing. Table 4.4 giving the income distribution of the families is interesting, though it was no doubt difficult for the informants to give the income of the members residing away in other income units. In para 4.41 it has been sought to be made out that the marriage rates among the category of daughters is a better indicator of the future pattern. But if ages at marriage are recorded, the marriage rates at the lower age groups should be an equally good indicator and the best will be the behaviour of recent marriage cohorts; one has not to go through the bio-family to get them.

The information on migration obtained from the bio-family unit is new and interesting; the value of the information would have been enhanced if there were no overlap in the unit and if distinction had been made between migration and just temporary visits. The information obtained on family planning is also very interesting, though it is not first of its kind in India as has been claimed in the foreword; objective data on family planning and attitudes were gathered by Dr. Chandrasekaran and Dr. Mukta Sen of the All India Institute of Hygiene and Public Health, Calcutta and tables based on them were published in the book by Davis on the '*Population of India and Pakistan*'. The data on family planning collected in the present survey are however broader based, though the high proportion of

## BOOK REVIEW

non-response is disturbing. It is significant that questions on family planning and attitudes offended a number of informants.

In concluding, the warning administered in para 5.11 of the report to all investigators against the belief that to obtain information on any point one has merely to frame a question and ask, could be profitably re-iterated. It is realised that some nods are inevitable in a work of this nature when the authors ventured in a rather unchartered field, and the criticisms are not meant to detract from the merit of their achievement.

Ajit Das Gupta

**A Study in the Analysis of Stationary Time Series :** By Herman Wold; Second Edition, with an appendix by Peter Whittle; Almquist and Wicksell, Stockholm; Price Sw. Kr. 28/-.

The publication of a second edition of Professor Wold's well known book will be welcomed by all mathematical statisticians, and especially by Time Series analysts who are all indebted to this book in a very large measure. For, this book is not only 'the first of its kind', as Neyman noted in 1939, it also happens to be the last. This is the only book on Time Series Analysis we have in statistical literature; and even 15 years after its first publication, it remains indispensable to all interested in the subject.

Professor Wold joins up the line of work of the English empiricists Arthur Schuster, Gilbert Walker, Udny Yule etc., with that of the Continental Probabilists, Kolmogoroff, Khintchine, etc. The first three chapters contain a penetrating analysis of the consequences of discrete time parameter stationarity as defined by Khintchine. All the different models that have been in use for describing observed oscillatory series are shown to be just particular cases of the Discrete Stationary Process. Professor Wold himself proves some interesting theorems, of which perhaps the most important is the Decomposition Theorem laying down a canonical form for the Discrete Stationary Process. This is an important result, mathematically elegant, and at the same time of great practical usefulness. In two other important theorems, he extends the notion of a spectrum to Discrete Processes and lays down the necessary and sufficient condition that  $k$  real numbers might be the first  $k$  autocorrelations of a moving average process.

The fourth chapter is devoted to applications of his theory to observational data. The method of estimation and of fitting models to data may not be regarded as very satisfactory from a modern stand point. Be as it may, the fact remains that these are problems to which satisfactory solutions have not as yet been provided by any body else.

Neyman, reviewing the first edition of this book, noted that there were several problems that were unsolved and hoped that they might be solved in future editions of the book. Many of these unsolved problems have been the subject of intense research since that time. But this second edition does not fulfill Neyman's hope. Except for minor alterations, the text of the present edition is unchanged. Attempt has been made to make it up-to-date by providing two appendices that replace those of the first edition. The first of these appen-

dices by Professor Wold himself consists of some notes referring to points in the main text. The second appendix by Peter Whittle purports to be a survey of the Spectral Theory and methods of Inference in Time Series. Unfortunately, it does not discuss any other methods of inference except those due to Dr. Whittle himself. Even Professor Wold's own goodness of fit test for the moving average is not even mentioned. Ably written as these two appendices are, they cannot provide satisfaction to those who would like to have a connected and systematic account of recent developments of the subject. We therefore hope that Professor Wold would bring out a new volume dealing in detail with such topics as the Spectral theory the concept of Ergodicity, the problems of estimation and testing of hypothesis in time series, etc.

A. Rudra



# SANKHYĀ

## THE INDIAN JOURNAL OF STATISTICS

*Edited by : P. C. MAHALANOBIS*

VOL. 17, PART 2

AUGUST

1956

### ON THE RECOVERY OF INTER BLOCK INFORMATION IN VARIETAL TRIALS

*By C. RADHAKRISHNA RAO*  
*Indian Statistical Institute, Calcutta*

#### 1. INTRODUCTION

The object of this note is to make further comments on the method of combined intra and inter block analysis of experiments suggested by the author earlier in 1947, derive explicit expressions in the case of linked block (LB) designs and suggest a new method more suitable to designs which admit easy estimation of block parameters as in the case of LB, lattice and similar designs (Roy and Laha, 1956; Ramakrishnan, 1956, appearing in this issue).

#### 2. THE Q METHOD

2.1. *The general analysis.* In the intra block analysis of varietal trials the normal equation for the estimation of varietal differences can be written, assuming equal replication for all varieties,

$$Q_i = \frac{r(k-1)}{k} t_i - \frac{\lambda_{i1}}{k} t_1 \dots - \frac{\lambda_{iv}}{k} t_v \quad \dots (2.1.1)$$

$$i = 1, \dots, v$$

together with the consistent equation

$$0 = t_1 + \dots + t_v$$

where  $Q_i$  = the total yield for the  $i$ -th variety minus the sum of block means in which it occurs,

$\lambda_{ij}$  = the number of blocks in which the  $i$ -th and  $j$ -th varieties occur together,

$r$  = the number of replications,

and  $k$  = the block size.



The corresponding equations for the combined intra and inter block analysis is shown to be (Rao, 1947, equation 3.12 with  $P$  changed to  $Q(c)$  where  $c$  stands for combined)

$$Q_i(c) = \frac{R(k-1)}{k} t_i - \frac{\Lambda_{i1}}{k} t_1 - \dots - \frac{\Lambda_{ir}}{k} t_r \quad \dots (2.1.2)$$

together with the consistent condition

$$0 = t_1 + \dots + t_r$$

where

$$R = r[w + w'/(k-1)],$$

$$\Lambda_{ij} = \lambda_{ij}(w - w'),$$

$$Q(c) = wQ_i + w'Q'_i \quad \dots (2.1.3)$$

$Q'_i$  = sum of means of blocks in which the  $i$ -th variety occurs minus  $r$  times the grand mean

$w$  = the reciprocal of the estimated intra block variance and  $w'$  that of inter block variance (Rao, 1947, equations 4.1, 4.2).

It is seen that the equality of any two  $\lambda_{ij}$  imply the equality of the corresponding  $\Lambda_{ij}$ . Since the equations (2.1.1) and (2.1.2) are similar, it was noted in Rao (1947) that the solutions of the former as functions of  $Q$ ,  $r$  and the distinct  $\lambda_{ij}$  provide solutions for the latter by writing  $Q(c)$ ,  $R$ ,  $\Lambda_{ij}$  for  $Q$ ,  $r$ ,  $\lambda_{ij}$ . The same is true of the expressions for the variances. It is necessary for the application of this method that  $r$  and  $\lambda_{ij}$  are treated as parameters and the solutions are obtained as their functions without recognizing the relationships such as

$$r(k-1) = \sum_i \lambda_{ij} \quad \dots (2.1.4)$$

or any other relationship among the  $\lambda_{ij}$  values, since such relationships may not be true in terms of  $R$  and  $\Lambda_{ij}$ . If a relationship such as (2.1.4) is explicitly used then  $R$  and  $\Lambda_{ij}$  should be defined as

$$R = r[w + w'(v-k)/v(k-1)]$$

$$\Lambda_{ij} = \lambda_{ij}(w - w') + \frac{rk}{v} w' \quad \dots (2.1.5)$$

With these new definitions it is immaterial whether  $r$  is considered as an independent parameter or not in solving the intra block equations.

It is, however, possible to solve the intra block equations by not recognizing any relationship among  $r$  and  $\lambda_{ij}$ , in which case the expressions for the combined case can be obtained by changes given in (2.1.3). A certain amount of care may be necessary involving the actual examination of the method of solution instead of depending on published formulae. It will be seen that the method of solving is similar for the equations (2.1.1) and (2.1.2) thus establishing the correspondence between the solutions. Let us consider a few examples.

# THE RECOVERY OF INTER BLOCK INFORMATION IN VARIETAL TRIALS

2.2. *The balanced incomplete block.* The intra block equations are

$$Q_i = \frac{r(k-1)}{k} t_i - \frac{\lambda}{k} \sum_{j \neq i} t_j, \quad i = 1, \dots, v \quad \dots (2.2.1)$$

$$0 = t_1 + \dots + t_v.$$

Eliminating  $\sum_{j \neq i} t_j$  by using the last equation

$$Q_i = \frac{r(k-1)+\lambda}{k} t_i$$

from which

$$t_i = \frac{kQ_i}{r(k-1)+\lambda}$$

and

$$V(t_i - t_j) = V \frac{k(Q_i - Q_j)}{r(k-1)+\lambda} = \frac{2k}{r(k-1)+\lambda} \sigma^2.$$

It is clear that, if instead of the equations (2.2.1) we had

$$Q(c) = \frac{R(k-1)}{k} t_i - \frac{k}{k} \sum_{j \neq i} t_j, \quad i = 1, \dots, v \quad \dots (2.2.2)$$

$$0 = t_1 + \dots + t_v.$$

We obtain the same expressions for estimates and variances with  $Q, r, \lambda$  changed to  $Q(c), R, \Lambda$ . In this special case the transformation is

$$\Lambda = \lambda(w - w')$$

$$R = r[w + w' / (k-1)].$$

For the variance,  $\sigma^2$  has to be dropped since it is already taken into account in the above transformation.

2.3. *Partially balanced incomplete block.* Let us consider a partially balanced design with two classes of associates (Bose and Nair, 1939; Nair and Rao, 1942). The  $Q$  equations are

$$Q_i = \frac{r(k-1)}{k} t_i - \frac{\lambda_1}{k} \sum_{1i} t_j - \frac{\lambda_2}{k} \sum_{2i} t_j, \quad i = 1, \dots, v \quad \dots (2.3.1)$$

$$0 = t_1 + \dots + t_v$$

where  $\sum_{1i}$  and  $\sum_{2i}$  indicate respectively the summations over the first and second associates of the  $i$ -th variety.

To solve these equations we follow the method followed by Bose and Nair (1939). By summing over the  $i$ -th equation and its first associates after eliminating  $\Sigma_{2i} t_j$  by using the last equation we find

$$\Sigma_{1i} Q_j = A_{22} t_i + B_{22} \Sigma_{1i} t_j$$

while the equation (2.3.1) can be written

$$Q_i = A_{12} t_i + B_{12} \Sigma_{1i} t_j$$

where

$$k A_{12} = r(k-1) + \lambda_2$$

$$k B_{12} = \lambda_2 - \lambda_1$$

$$k A_{22} = (\lambda_2 - \lambda_1) p_{12}^2$$

$$k B_{22} = r(k-1) + \lambda_2 + (\lambda_2 - \lambda_1)(p_{11}^1 - p_{11}^2).$$

Eliminating  $\Sigma_{1i} t_j$  and solving for  $t_i$  we get

$$t_i = [Q_i B_{22} + B_{12} \Sigma_{1i} Q_j] \div \Delta$$

$$= [(B_{22} + B_{12}) Q_i + B_{12} \Sigma_{2i} Q_j] \div \Delta$$

where

$$\Delta = A_{12} B_{22} - A_{22} B_{12}.$$

The second formula is convenient if the number of second associates is smaller in number. The variance of  $t_i - t_r$  is

$$2(B_{22} + B_{12})\sigma^2/\Delta$$

if  $t_i$  and  $t_r$  are first associates and

$$2B_{22}\sigma^2/\Delta$$

if they are second associates.

In this method the solution for varietal differences and the expressions for the variances would have been the same if  $Q(c)$ ,  $R$ ,  $\Lambda_1$ ,  $\Lambda_2$  as defined in (2.1.3) were used instead of  $Q$ ,  $r$ ,  $\lambda_1$ ,  $\lambda_2$ . Hence the expressions for the combined analysis can be obtained from the above by changing  $Q$ ,  $r$ ,  $\lambda$  to  $Q(c)$ ,  $R$ ,  $\Lambda$ . In the expressions for the variance,  $\sigma^2$  should be dropped.

The analysis is similar for designs involving more associates. Special cases such as lattice designs, triangular designs etc. may be considered in a similar way and explicit expressions obtained or the general expression derived above may be used with special values of  $\lambda_1$ ,  $\lambda_2$  and  $p_{jk}^i$ . The general expressions for designs with three associate classes are given in Rao (1947) in such way as to provide combined estimates by changing  $Q$ ,  $r$ ,  $\lambda$  to  $Q(c)$ ,  $R$ ,  $\Lambda$ .

2.4. *Linked blocks.* In linked block (LB) designs introduced by Youden (1951) there are  $v$  varieties each replicated  $r$  times,  $b$  blocks of  $k$  plots and any two blocks have  $\mu$  varieties in common. Roy and Laha (1956) provided a very simple

THE RECOVERY OF INTER BLOCK INFORMATION IN VARIETAL TRIALS

method of analysing such designs using only the linked block property. This method will be explained in § 3 (the  $P$  method) where an alternative approach to intra and inter block analysis is considered.

For convenience in proving some results, we introduce matrix notation and express the equations in terms of matrices. Let  $N$  be the incidence matrix with  $b$  rows and  $v$  columns representing the blocks and varieties. The  $(i, j)$ -th element is unity if the  $i$ -th block contains the  $j$ -th variety and zero otherwise. It is easy to see for LB designs

$$NN' = \begin{pmatrix} k & \mu & \dots & \mu \\ . & . & \dots & . \\ \mu & \mu & \dots & k \end{pmatrix} = (k-\mu)I + \mu U$$

where  $U$  is a matrix with all elements unity and  $I$  is the unit matrix. Also

$$N'N = \begin{pmatrix} r & \lambda_{12} & \dots & \lambda_{1v} \\ . & . & \dots & . \\ \lambda_{v1} & \lambda_{v2} & \dots & r \end{pmatrix}$$

where  $\lambda_{ij}$  is the number of blocks in which the  $i$ -th and  $j$ -th varieties occur together. The intra-block  $Q$  equations are

$$Q = \left( rI - \frac{1}{k} N'N \right) \underline{t} \quad \dots (2.4.1)$$

where  $Q$  is the column vector  $(Q_1, \dots, Q_v)$  and  $\underline{t}$ , the column vector  $(t_1, \dots, t_v)$ .

Multiplying both sides by  $N'N$  and simplifying

$$\begin{aligned} N'N Q &= \left( rN'N - \frac{1}{k} N'NN'N \right) \underline{t} \\ &= \left( rN'N - \frac{k-\mu}{k} N'N - \frac{\mu}{k} N'UN \right) \underline{t} \\ &= \left( r - \frac{k-\mu}{k} \right) N'N \underline{t} \end{aligned} \quad \dots (2.4.2)$$

since  $N'UN \underline{t} = 0$ .

Eliminating  $N'N \underline{t}$  from (1.4.1) and (2.4.2)

$$\begin{aligned} [r(k-1)+\mu]Q + N'N Q &= r[r(k-1)+\mu]\underline{t} \\ \underline{Q} + \frac{N'N Q}{r[r(k-1)+\mu]} &= \underline{t} \end{aligned} \quad \dots (2.4.3)$$



Comparing elements on both sides

$$\begin{aligned} t_i &= \frac{Q_i}{r} + \frac{rQ_i + \sum \lambda_{is} Q_s}{r[r(k-1) + \mu]} \\ &= Q_i \left( \frac{1}{r} + \frac{1}{\mu b} \right) + \frac{\sum \lambda_{is} Q_s}{\mu b r}, \end{aligned}$$

since

$$r(k-1) = \mu(b-1).$$

The variance of  $t_i - t_j$  is

$$2 \left\{ \left( \frac{1}{r} + \frac{1}{\mu b} \right) - \frac{\lambda_{ij}}{\mu b r} \right\} \sigma^2$$

which depends only on  $\lambda_{ij}$  besides the parameters  $r, b, \mu$  as shown by Roy and Laha (1956).

From the above analysis it is not clear how the combined intra and inter block estimates can be obtained. We shall write down the combined equations and follow the above procedure. Using matrix notation the combined equations

$$Q_i(c) = \frac{R(k-1)}{k} t_i - \frac{\Lambda_{ii}}{k} t_1 \cdots - \frac{\Lambda_{ir}}{k} t_r$$

of (2.1.2) can be written

$$\underline{Q}(c) = \left( rw I - \frac{w-w'}{k} N'N \right) \underline{t}. \quad \dots (2.4.4)$$

We follow the same procedure as above. Multiplying both sides by  $N'N$  and simplifying we get

$$\begin{aligned} N'N \underline{Q}(c) &= \left[ rw N'N - \frac{(w-w')(k-\mu)}{k} N'N \right] \underline{t} \\ &= \left[ rw - \frac{w-w'}{k} (k-\mu) \right] N'N \underline{t} = g(w-w') N'N \underline{t} \end{aligned}$$

where

$$g = [krw - (k-r)(w-w')]/(w-w'). \quad \dots (2.4.5)$$

Eliminating  $N'NT$  from (2.4.4) and (2.4.5) as before

$$g \underline{Q}(c) + N'N \underline{Q}(c) = grw \underline{t}. \quad \dots (2.4.6)$$

Comparing the elements in the vectors on both sides

$$t_i = \frac{(g+r)Q_i(c) + \sum \lambda_{is} Q_s(c)}{grw}.$$

# THE RECOVERY OF INTER BLOCK INFORMATION IN VARIETAL TRIALS

The variance of  $t_i - t_j$  is

$$2 \frac{g+r-\lambda_{ij}}{grw} \quad \dots (2.4.7)$$

which depends only on  $\lambda_{ij}$  besides other parameters common to all pairs  $(i, j)$ .

We thus obtain the expressions for the combined analysis by following the same method of solving although it may be difficult to obtain the intra-block solutions in such a way as to provide the combined solutions by changing  $Q, r, \lambda$  to  $Q(c), R$  and  $\Lambda$ . But this is clearly unnecessary unless it is simpler to do so.

## 3. THE P METHOD

3.1. *Intra block analysis.* Adopting matrix notation as in sub-section 2.4, we write the  $Q$  equations

$$\underline{Q} = \left( rI - \frac{1}{k} N'N \right) \underline{t}. \quad \dots (3.1.1)$$

Instead of solving these equations directly we may add to it  $b$  other consistent equations

$$\underline{B} = k\underline{b} + N\underline{t} \quad \dots (3.1.2)$$

where  $\underline{B}$  = the column vector containing the block totals

$\underline{b}$  = the column vector of  $b$  additional constants  $b_1, b_2, \dots, b_b$  which may be referred to as block constants.

Multiplying (3.1.2) by  $\frac{1}{k} N'$  and adding to (3.1.1) gives

$$\underline{T} = r\underline{t} + N'\underline{b} \quad \dots (3.1.3)$$

where  $\underline{T}$  is the column vector of total yields of varieties.

Eliminating  $\underline{t}$  from the equations (3.1.2), (3.1.3)

$$\underline{P} = \underline{B} - \frac{1}{r} N\underline{T} = \left( k - \frac{NN'}{r} \right) \underline{b}. \quad \dots (3.1.4)$$

Writing these equations in full

$$P_j = \frac{k(r-1)}{r} b_j - \frac{\mu_{j1}}{r} b_1 - \dots - \frac{\mu_{jb}}{r} b_b \quad \dots (3.1.5)$$

where  $P_j$  = total yield of the  $j$ -th block minus the sum of the mean yields of varieties occurring in the  $j$ -th block.

$\mu_{ij}$  = the number of varieties common to the  $i$ -th and  $j$ -th blocks.

The equations (3.1.5) will be referred to as the  $P$  equations. It may be easier to solve these equations for the  $b$  constants which may be subject to a restriction of the type

$$b_1 + \dots + b_b = 0.$$

Having obtained these values they may be substituted in (3.1.3) to obtain  $t_i$ . The  $i$ -th equation in (3.1.3) is

$$T_i = r t_i + \Sigma_i b_s$$

where  $\Sigma_i$  denotes summation over the blocks in which the  $i$ -th variety occurs. Hence

$$t_i = \frac{1}{r} T_i - \frac{1}{r} \Sigma_i b_s.$$

The estimate of  $t_i - t_j$  is  $\frac{1}{r} \{T_i - T_j - (\Sigma_i b_s - \Sigma_j b_s)\}$  and the variance of  $(t_i - t_j)$  is

$$\begin{aligned} V(t_i - t_j) &= \frac{1}{r^2} V(T_i - T_j) + \frac{1}{r^2} V(\Sigma_i b_s - \Sigma_j b_s) \\ &= \sigma^2 \left\{ \frac{2}{r} + \frac{C_{ij}}{r^2} \right\} \end{aligned}$$

where  $\sigma^2 C_{ij} = V(\Sigma_i b_s - \Sigma_j b_s)$ .

To compute this variance the simplest way is to find the expression for  $\Sigma_i b_s - \Sigma_j b_s$  in terms of  $P_j$  and use the following general result.

If 
$$l_1 b_1 + \dots + l_b b_b = d_1 P_1 + \dots + d_b P_b$$

then 
$$V(l_1 b_1 + \dots + l_b b_b) = (d_1 l_1 + \dots + d_b l_b) \sigma^2.$$

The sum of squares due to varieties can also be obtained by using the solutions of the  $P$  equations as was first shown by Yates (1939, 1940). So the entire analysis can be carried out by using the  $P$  equations (3.1.5) and the  $T$  equations (3.1.3) involving varietal totals. This may be referred to as the  $P$  method. In the next section it is shown that the  $P$  method could be used in the intra and inter block analysis also.

3.2. *The intra and inter block analysis.* To the  $Q$  equations (2.1.2) for the combined intra and inter block estimates

$$Q(c) = \left( w r I - \frac{w - w'}{k} N' N \right) t \quad \dots \quad (3.2.1)$$

we add the consistent equations

$$B - km\bar{u} = k\bar{b} + N\bar{t} \quad \dots (3.2.2)$$

where  $\bar{u}$  is the column vector with all its elements unity and  $m$ , the grand mean.

Eliminating  $N'N\bar{t}$  as before

$$Q(c) + \frac{w-w'}{k} N'(B - km\bar{u}) = wr\bar{t} + (w-w')N'\bar{b}$$

or writing out in full

$$w(T_i - rm) = wrt_i + (w-w')\Sigma_i b_s, \quad i = 1, \dots, v. \quad \dots (3.2.3)$$

Substituting for  $t_i$  in (3.2.2) we obtain

$$w P_i = b_i k \left( w - \frac{w-w'}{r} \right) - \frac{w-w'}{r} \Sigma \mu_{is} b_s \quad \dots (3.2.4)$$

where  $P_i$  is same as defined in (3.1.5). The equations (3.2.4) are solved in the same way as (3.1.5). The values of  $b$  are substituted in (3.2.3) to obtain a solution for  $t_i$

$$wrt_i = wT_i - (w-w')\Sigma_i b_s$$

$$t_i - t_j = \frac{w(T_i - T_j)}{wr} - \frac{(w-w')(\Sigma_i b_s - \Sigma_j b_s)}{wr}.$$

If

$$\Sigma_i b_s - \Sigma_j b_s = w(c_1 P_1 + c_2 P_2 + \dots + c_b P_b)$$

then

$$V(t_i - t_j) = \frac{2}{rw} \left\{ 1 + \frac{(w-w')}{r} (\Sigma_i c_s - \Sigma_j c_s) \right\}.$$

As an application let us consider the analysis of the LB design. The  $P$  equations (3.2.4) for combined estimation are

$$\begin{aligned} w P_i &= b_i k \left( w - \frac{w-w'}{r} \right) - \mu \frac{(w-w')}{r} \Sigma b_s \\ &= b_i \left\{ kw - \frac{(k-\mu)(w-w')}{r} \right\} - \mu(b_1 + \dots + b_b). \end{aligned}$$

Setting  $(b_1 + \dots + b_b = 0)$  the solutions are

$$\begin{aligned} b_i &= rw P_i / [r k w - (k-\mu)(w-w')] \\ &= rw P_i / g(w-w') \end{aligned}$$



where  $g$  is as defined in (2.4.6)

$$g(w-w') = k r w - (k-\mu)(w-w').$$

Substituting for  $b_i$  in 3.2.3

$$t_i = \frac{T_i}{r} - \frac{1}{g} \sum_i P_s$$

omitting the constant  $m$  since only the differences of  $t_1, \dots, t_r$  are to be considered

$$(t_i - t_j) = \frac{1}{r} (T_i - T_j) - \frac{1}{g} (\sum_i P_s - \sum_j P_s)$$

$$V(t_i - t_j) = \frac{2}{w} \left\{ \frac{1}{r} + \frac{1}{g} - \frac{\lambda_{ij}}{rg} \right\}$$

which is same as the expression (2.4.7) obtained by following the  $Q$  method.

#### REFERENCES

- BOSE, R. C. AND NAIR, K. R. (1939): Partially balanced incomplete block designs. *Sankhyā*, **4**, 237.  
 NAIR, K. R. AND RAO, C. R. (1942): A note on partially balanced incomplete block designs. *Science and Culture*, **7**, 516.  
 RAMAKRISHNAN, C. S. (1956): On the dual of a PBIB design, and a new class of designs with two replications. *Sankhyā*, **17**, 133-142.  
 RAO, C. R. (1947): General methods of analysis for incomplete block designs. *J. Amer. Stat. Ass.*, **42**, 541.  
 ROY, J. AND LAHA, R. G. (1956): Classification and analysis of linked block designs. *Sankhyā*, **17** 115-132.  
 YATES, F. (1939): The recovery of inter-block information in varietal trials arranged in three dimensional lattice. *Ann. Eugenics*, **9**, 136.  
 ——— (1940): The recovery of inter-block information in balanced incomplete block designs. *Ann. Eugenics*, **10**, 317.

*Paper received : February, 1956.*

# CLASSIFICATION AND ANALYSIS OF LINKED BLOCK DESIGNS

By J. ROY and R. G. LAHA

*Indian Statistical Institute, Calcutta*

## 1. INTRODUCTION AND SUMMARY

In comparative experiments involving a fairly large number of varieties, when for lack of homogeneous experimental units, complete block designs are not available Balanced Incomplete Block (BIB) designs prove very convenient in two respects. Firstly, the analysis is very simple and secondly, comparison between any two varieties has the same precision. On the other hand, a BIB design requires a large number of experimental units, because balancing is not possible unless the number of blocks is atleast as large as the number of varieties. To obviate this difficulty, different types of incomplete block designs have been introduced, of which the most notable is the Partially Balanced Incomplete Block (PBIB) design introduced by Bose and Nair (1939). Another class of incomplete block designs, called Linked Block (LB) designs was introduced by Youden (1951) which he obtained by dualising several BIB designs, that is by taking the varieties and blocks of the BIB design respectively as blocks and varieties in the LB design. The LB designs so constructed by Youden happen to be all PBIB designs, but this is not necessarily true. A great advantage of LB designs is that the analysis can be easily worked out.

In this paper, we show how the intra-block analysis of a LB design can be neatly carried out. The efficiency factor of a LB design is obtained. The methods are illustrated with a numerical example. An exhaustive list of all LB designs with ten or less plots per block and involving ten or less replications is given. These designs fall into three groups: (1) symmetrical BIB designs, (2) PBIB designs and (3) Irregular designs. For plans of LB designs belonging to group (2) reference is made to the serial number of the two-associate class PBIB designs enumerated by Bose, Clatworthy and Shrikhande (1954). Plans for other designs belonging to groups (2) and (3) are given in detail. Certain theorems are derived which are useful in determining from the parameters of a given two-associate class PBIB design whether the design is of the LB type or not.

## 2. SOME DEFINITIONS

An arrangement of  $v$  varieties in  $b$  blocks, each of  $k$  plots,  $k < v$  such that each variety occurs atmost once in any block and altogether in  $r$  blocks will be called and equi-replicate incomplete block design. Such a design is completely characterized by its "incidence-matrix"

$$N \equiv ((n_{ij}))$$

where  $n_{ij} = 1$  if the  $j$ -th variety occurs in the  $i$ -th block  
and 0 otherwise  $i = 1, 2, \dots, b; j = 1, 2, \dots, v$ .

Let  $\lambda_{ij}$  denote the number of blocks in which the  $i$ -th and the  $j$ -th varieties occur together  $i \neq j = 1, 2, \dots, v$  and  $\mu_{ij}$  the number of varieties which occur both in the  $i$ -th and  $j$ -th blocks  $i \neq j = 1, 2, \dots, b$ .

Then

$$\lambda_{ij} = \sum_{t=1}^b n_{it}n_{jt}$$

$$\mu_{ij} = \sum_{u=1}^v n_{iu}n_{ju}$$

For the sake of completeness write

$$\lambda_{ii} = \sum_{t=1}^b n_{it}^2 = r$$

$$\mu_{ii} = \sum_{u=1}^v n_{iu}^2 = k$$

We shall call the matrix  $\Lambda \equiv ((\lambda_{ij}))$  the 'association matrix' of the design and  $M \equiv ((\mu_{ij}))$  the 'block-characteristic matrix' of the design. Obviously  $\Lambda = N'N$  and  $M = NN'$ .

An equi-replicate incomplete block design is called a Linked Block (LB) design if  $\mu_{ij} = \mu$  for all  $i \neq j = 1, 2, \dots, b$  and it is said to be a Balanced Incomplete Block (BIB) design if  $\lambda_{ij} = \lambda$  for all  $i \neq j = 1, 2, \dots, v$ . It is well known that the necessary and sufficient condition for a BIB design to be a LB design and vice-versa is that  $v = b$ .

A design obtained from a given design by considering its blocks as varieties and varieties as blocks is said to be its dual. Obviously a LB design is the dual of some BIB design and vice-versa. Since a BIB design requires at least as many blocks as the number of varieties in a LB design the number of blocks can not exceed the number of varieties.

The following definition of a PBIB design from Bose and Shimamoto (1952) will be required in later sections:

A design is said to be a PBIB design with  $m$  associate classes if there exists a relationship of association between every pair of the  $v$  varieties satisfying the following conditions:

(a) Any two varieties are either first, second, ..., or  $m$ -th associates and any pair of varieties which are  $s$ -th associates occur together in  $\lambda_s$  blocks, ( $s = 1, 2, \dots, m$ ).

(b) Each variety has  $n_s$   $s$ -th associates

(c) For any pair of varieties which are  $s$ -th associates, the number of varieties which are simultaneously the  $j$ -th associates of the first and  $u$ -th associates of

# CLASSIFICATION AND ANALYSIS OF LINKED BLOCK DESIGNS

the second is  $p_{ju}^s$  and this is independent of the pair of varieties with which we start. Furthermore

$$p_{ju}^s = p_{uj}^s (j \neq u, j, s, u, = 1, 2, \dots, m).$$

It is known that the following conditions are satisfied by the parameters:

$$\sum_{s=1}^m n_s = v-1$$

$$\sum_{s=1}^m n_s \lambda_s = r(k-1)$$

$$\sum_{s=1}^m p_{js}^i = \begin{cases} n_i - 1 & \text{if } i = j \\ n_i & \text{if } i \neq j \end{cases}$$

$$n_i p_{ju}^i = n_j p_{iu}^j = n_u p_{ij}^u.$$

## 3. INTRA-BLOCK ANALYSIS OF LINKED BLOCK DESIGNS

Consider a Linked Block (LB) design involving  $v$  varieties in  $b$  blocks, each of  $k$  plots, each variety replicated  $r$  times in which any two blocks have  $\mu$  varieties in common. Obviously,

$$bk = rv, \quad \mu(b-1) = k(r-1), \quad b \leq v.$$

Let  $N \equiv ((n_{ij}))$  be the incidence matrix of this design.

$$\text{Then} \quad \sum_{i=1}^b n_{ij} = r, \quad \sum_{j=1}^v n_{ij} = k, \quad \sum_{j=1}^v n_{ij} n_{i'j} = \mu.$$

Let  $y_{ij}$  denote the yield from that plot of the  $i$ -th block which gets the  $j$ -th treatment. Then, denoting the effect of the  $i$ -th block by  $\beta_i$  and that of the  $j$ -th treatment by  $\tau_j$  the normal equations for estimating these parameters from intra-block differences (under the usual assumption of additivity of block and treatment effects) are:

$$T_j = r \tau_j + \sum_{i=1}^b n_{ij} \beta_i \quad (j = 1, 2, \dots, v) \quad \dots (3.1)$$

$$B_i = k \beta_i + \sum_{j=1}^v n_{ij} \tau_j \quad (i = 1, 2, \dots, b) \quad \dots (3.2)$$

where  $T_j$  stands for the total yield of all plots getting the  $j$ -th variety and  $B_i$  that for all plots in the  $i$ -th block.



Eliminating  $\tau_j$ 's from (3.1) and (3.2) we get

$$\hat{\beta}_i = \frac{r}{\mu b} P_i + \bar{\beta}$$

and finally

$$\hat{\tau}_j = t_j - \bar{\beta}$$

where

$$P_i = B_i - \frac{1}{r} \sum_{j=1}^v n_{ij} T_j. \quad \dots (3.3)$$

$$t_j = \frac{1}{r} T_j - \frac{1}{\mu b} \sum_{i=1}^r n_{ij} P_i. \quad \dots (3.4)$$

Thus the parameters  $\beta_i$  and  $\tau_j$  are estimable except for an additive indeterminate  $\bar{\beta}$ .

Let us take

$$t_j$$

and

$$b_i = \frac{1}{\mu b} P_i \quad \dots (3.5)$$

for a particular solution of the normal equations (3.1) and (3.2). Then any varietal contrast  $\sum_{j=1}^v l_j \tau_j$  with  $\sum_{j=1}^r l_j = 0$  is estimable and its best intra-block linear estimate is  $\sum_{j=1}^v l_j t_j$  with variance given by

$$\sigma^2 \sum_{j,j'=1}^v l_j l_{j'} C_{jj'}, \quad \dots (3.6)$$

where

$$C_{jj} = \frac{1}{r} + \frac{1}{\mu b}$$

and

$$C_{jj'} = \frac{\lambda_{jj'}}{r \mu b} \quad \dots (3.7)$$

$\lambda_{jj'}$  denoting the number of blocks in which the  $j$ -th and  $j'$ -th varieties occur together and  $\sigma^2$  is the intra-block error variance which can be estimated from the error component in the analysis of variance given below:

TABLE 1. ANALYSIS OF VARIANCE

variation due to	S.S.	d.f.	S.S.	variation due to
blocks (unadjusted)	$S_B^* = \frac{1}{k} \sum B_i^2 - CF$	$b-1$	$\frac{r}{\mu b} \sum P_i^2 = S_B$	blocks (adjusted)
varieties (adjusted)	$S_V = S_B - S_B^* - S_E^*$	$v-1$	$\frac{1}{r} \sum T_j^2 - CF = S_V^*$	varieties (unadjusted)
error	$S_E = S - S_V - S_B^*$	$bk-b-v+1$	$S - S_B - S_V^* = S_E$	error
total	$S = \sum \sum y_{ij}^2 - CF$	$bk-1$	$\sum \sum y_{ij}^2 - CF$	total
Grand total = G, $CF = G^2/bk$ .				

# CLASSIFICATION AND ANALYSIS OF LINKED BLOCK DESIGNS

We have seen that the variance of the best estimate of a varietal difference, say,  $\tau_j - \tau'_j$  is

$$2\sigma^2 \left\{ \frac{1}{r} + \frac{1}{\mu b} - \frac{\lambda_{jj'}}{r\mu b} \right\}$$

so that the average variance of the estimate of a treatment difference is

$$\frac{2\sigma^2}{r} \left\{ 1 + \frac{r - \bar{\lambda}}{\mu b} \right\}$$

$$\bar{\lambda} = r(k-1)/(v-1).$$

where

In a randomised block design with the same intra-block error variance and the same number of replications, the variance of the estimate of any varietal difference is  $2\sigma^2/r$ .

This bears to the former the ratio

$$E = 1 / \left\{ 1 + \frac{r - \bar{\lambda}}{\mu b} \right\} \quad \dots (3.8)$$

which is defined to be the efficiency of the design. Since  $r \geq \bar{\lambda}$ ,  $E \leq 1$  and  $E = 1$  implies  $v = k$ .

## 4. NUMERICAL ILLUSTRATION

We give below the computational details of the intra-block analysis of a LB design involving 18 varieties in 9 blocks each containing 8 plots, in which each variety is replicated 4 times and the number of varieties common to any two blocks is 3. This is design number 24 in our list given in § 7. The peculiarity of this design is that pairs of varieties may occur together in 0, 1, 2 or 3 blocks, but it does not belong to any of the well known designs—e.g. the PBIB designs.

The plan and the yields are given in Table 2. Varieties are indicated by numbers in brackets and corresponding yields are written by their sides. In Table 3, the numerical details of estimating varietal effects are given and in Table 4 the intra-block analysis of variance is presented. Table 5 gives the estimated standard errors of estimates of different types of varietal differences.

TABLE 2. PLAN AND YIELD

blocks		numbers of varieties and corresponding yields							
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	
1	( 7) 57	( 8) 53	( 5) 54	( 4) 58	( 3) 51	( 1) 54	( 2) 53	( 6) 48	
2	( 9) 52	(13) 53	(10) 54	(11) 56	( 2) 56	(12) 60	( 1) 56	( 3) 48	
3	(12) 65	(17) 57	(14) 66	( 4) 62	( 9) 52	( 6) 53	(16) 66	( 1) 60	
4	(16) 63	(10) 60	( 7) 62	( 5) 57	( 1) 59	(18) 58	(15) 57	(12) 65	
5	( 4) 63	( 2) 60	(13) 56	(10) 57	( 8) 54	(16) 61	(14) 69	(18) 58	
6	(16) 59	(17) 55	( 6) 47	(11) 52	(13) 47	(15) 54	( 5) 58	( 2) 53	
7	(18) 54	(15) 55	(17) 52	( 3) 51	( 8) 51	(11) 54	( 4) 56	(12) 59	
8	(13) 52	( 7) 62	( 9) 60	(14) 57	(18) 53	(17) 58	( 5) 60	( 3) 51	
9	(10) 46	(15) 51	(14) 54	(11) 50	( 8) 48	( 7) 62	( 6) 39	( 9) 54	

TABLE 3. ESTIMATION OF VARIETAL EFFECTS

$j$	$T_j$	$\sum_i n_{ij}(rP_i)$	$\mu br t_j$	$t_j$	$i$	$B_i$	$\sum_j n_{ij}T_j$	$rP_i$			
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)			
1	229	93	6090	56.39	1	428	1756	-44			
2	222	-4	5998	55.54	2	435	1756	-16			
3	201	-79	5506	50.98	3	481	1839	85			
4	239	102	6351	58.81	4	481	1856	68			
5	229	0	6183	57.25	5	478	1810	102			
6	187	-135	5184	48.00	6	425	1746	-46			
7	243	-84	6645	61.53	7	432	1769	-41			
8	206	-113	5675	52.55	8	453	1790	22			
9	218	-39	5925	54.86	9	404	1746	-130			
10	217	24	5835	54.03	total	4017	16068	0			
11	212	-233	5957	55.16							
12	249	96	6627	61.36	Computational checks:						
13	208	62	5554	51.43							
14	246	79	6563	60.77							
15	217	-149	6008	55.63							
16	249	209	6514	60.31							
17	222	20	5974	55.31							
18	223	151	5870	54.35							
total	4017	0	108459	1004.26							

$$\sum_i (rP_i) = 0$$

$$\sum_j \sum_i n_{ij}(rP_i) = 0$$

$$r \sum t_j = G$$

Here  $v = 18$ ,  $b = 9$ ,  $r = 4$ ,  $k = 8$ ,  $\mu = 3$

Grand total  $G = 4017$ ,

$$CF = G^2/bk = 224115.12$$

Total S.S.  $= S = \sum y^2 - CF$

$$= 2019.88$$

Block S.S. (unadjusted)  $= S_B^* = \frac{1}{k} \sum B_i^2 - CF$

$$= 796.00$$

Variety S.S. (unadjusted)  $= S_V^* = \frac{1}{r} \sum T_j^2 - CF$

$$= 1280.63$$

Block S.S. (adjusted)  $= S_B = \frac{1}{\mu br} \sum (rP_i)^2$

$$= 422.46$$

Variety S.S. (adjusted)  $= S_V = S_V^* + S_B - S_B^*$

$$= 907.09$$

Error S. S.  $= S_E = S - S_V^* - S_B = S - S_B - S_V^*$

$$= 316.79$$

# CLASSIFICATION AND ANALYSIS OF LINKED BLOCK DESIGNS

TABLE 4. ANALYSIS OF VARIANCE

variation due to	d.f.	s.s.	m.s.	F	m.s.	s.s.	d.f.	variation due to
blocks (unadjusted)	8	796.00	—	7.66	52.81	422.46	8	blocks (adjusted)
varieties (adjusted)	17	907.09	53.36	7.74	—	1280.63	17	varieties (unadjusted)
error	46	316.79	6.89		6.89	316.79	46	error
total	71	2019.88				2019.88		total

TABLE 5. STANDARD ERROR OF VARIETAL DIFFERENCES

number of blocks in which a given pair of varieties occur together	$\lambda$	$f_{\lambda} = \frac{2}{r} \left\{ 1 + \frac{r-\lambda}{\mu b} \right\}$	standard error of varietal difference $\left\{ \frac{f_{\lambda} \cdot S_E}{bk-b-v+1} \right\}^{\frac{1}{2}}$
	(1)	(2)	(3)
	0	0.574074	1.988
	1	0.555556	1.956
	2	0.537037	1.923
	3	0.518519	1.890

Table 5. has to be used in carrying out a  $t$ -test to examine any particular varietal difference. For example, if we have to see if varieties 1 and 7 are different, we observe that they occur together in  $\lambda = 2$  blocks. The estimate of the difference  $\tau_1 - \tau_7$  is  $t_1 - t_7 = 56.39 - 61.53 = -5.14$  obtained from Table 3 and its standard error is 1.923 as given in Table 5. Hence the statistic to use is

$$t = -5.14/1.923 = -2.673$$

which as a  $t$ -statistic with 46 degrees of freedom is significant at 1% level.

## 5 CLASSIFICATION OF LINKED BLOCK DESIGNS

LB designs are duals of BIB designs and Shrikhande (1952) has shown that the dual of a BIB design is a PBIB design with two associate classes in the following cases:

- (a) if in the BIB design  $\lambda = 1$  ( $v \neq b$ )
- (b) if in the BIB design  $\lambda = 2$ ,  $r = k+2$ ,  $k = k$ .

Roy (1954) has shown further that the dual of any unreduced BIB design (that is a design obtained by forming a block with each possible combination of varieties subject to the condition that the number of plots in a block is fixed) is necessarily of the



PBIB type. Again, if a BIB design is symmetric, that is if in a BIB design the number of blocks is equal to the number of varieties, then its dual is also a BIB design. However, the dual of any BIB design is not necessarily a BIB or even PBIB design.

We may therefore classify all LB designs into the following three main groups: (1) Symmetrical BIB designs (2) PBIB designs and (3) Irregular designs not belonging to any of the known types. We shall omit the symmetrical BIB designs from our consideration as they do not present any new features.

In order to get the LB design of the PBIB type, one way is to start with the plans of all BIB designs, dualise them and then pick out from them the designs of the PBIB type. This, however, is a formidable task. On the other hand, Bose, Clatworthy and Shrikhande (1954) have prepared a list of useful two associate PBIB designs and enumerated their plans and parameters. It is considerably simpler to pick out from this list the LB designs with the use of the theorems derived in the next section which state the necessary and sufficient conditions on the parameters of a PBIB design so that it may be of the LB type.

In § 7 we give a detailed classified list of all LB designs with ten or less replications and ten or less plots per block.

#### 6. CONDITION THAT A TWO-ASSOCIATE PBIB DESIGN MAY BE OF THE LB TYPE

It is a well known property of matrices that given any two matrices  $A$ ,  $B$  such that the matrix products  $AB$  and  $BA$  are possible, the non-zero roots of the two determinantal equations

$$|AB - xI| = 0$$

$$|BA - xI| = 0$$

are identical. It follows therefore that for any design, the association matrix and the block characteristic matrix have the same set of non-zero latent roots.

Now the block-characteristic matrix of a LB design has all diagonal elements equal to  $k$  and all off-diagonal elements equal to  $\mu$ ,  $k$  being the number of plots per block and  $\mu$  the number of varieties common to any two blocks. Consequently the block-characteristic matrix of a LB design has only two distinct latent roots, namely  $k + (b-1)\mu = rk$  and  $k - \mu$  the latter of multiplicity  $(b-1)$ . On the other hand, for the block characteristic matrix, if  $rk$  is a latent root with  $(1, 1, \dots, 1)$  for latent vector, and  $k - \mu$  is another latent root of multiplicity  $(b-1)$ , all other roots being zero, the design must be of the LB type. Again, it follows from the work of Connor and Clatworthy (1954) that latent roots other than  $rk$  of the association matrix of a two-associate PBIB design are except for repetitions the same as those of the matrix

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

# CLASSIFICATION AND ANALYSIS OF LINKED BLOCK DESIGNS

where

$$\left. \begin{aligned} a_{11} &= r + \lambda_1 p_{11}^1 + \lambda_2 p_{12}^1 - \lambda_1 n_1 \\ a_{12} &= \lambda_1 p_{11}^2 + \lambda_2 p_{12}^2 - \lambda_1 n_1 \\ a_{21} &= \lambda_1 p_{21}^1 + \lambda_2 p_{22}^1 - \lambda_2 n_2 \\ a_{22} &= r + \lambda_1 p_{21}^2 + \lambda_2 p_{22}^2 - \lambda_2 n_2 \end{aligned} \right\} \dots \quad (6.1)$$

where the parameters have their usual significance. They have also shown that the matrix  $A$  can not have two equal latent roots. Hence we have the following

**Theorem 6.1.** *The necessary and sufficient condition for a PBIB design with two associate classes to be of the LB type is that the matrix  $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$  defined in (6.1) has only one non-zero root and this occurs  $(b-1)$  times as a latent root of the association matrix of the PBIB design.*

Two associate PBIB designs have been classified by Bose, Clatworthy and Shrikhande (1954) as (1) Group Divisible (a) Singular (b) Semi-Regular (c) Regular (2) Triangular (3) Latin Square Type (4) Simple and (5) Cyclic. Theorem 6.1 gives the following conditions on the parameters of the first four of these designs which ensure that they belong to the LB type.

For a Group Divisible (GD) design

$$v = mn, n_1 = n-1, n_2 = m(n-1), p_{11}^1 = n-2 \text{ and } r \geq \lambda_1, rk \geq v\lambda_2.$$

A GD design is singular if  $r = \lambda_1$ , semi-regular if  $r > \lambda_1$ ,  $rk = v\lambda_2$  and regular if  $r > \lambda_1$ ,  $rk > v\lambda_2$ . For a singular GD design  $b \geq m$  and for a semi-regular GD design  $b \geq v-m+1$ . These results are due to Bose and Connor (1952). Application of theorem 6.1 gives the following result. A regular GD is never of the LB type, the necessary and sufficient condition for a singular GD to be of the LB type is that  $b = m$  and that for a semi-regular GD is that  $b = v-m+1$ .

In a Triangular design as defined by Bose and Shimamoto (1952)  $v = \frac{1}{2}n(n-1)$ ,  $n_1 = 2(n-2)$ ,  $n_2 = \frac{1}{2}(n-2)(n-3)$ ,  $p_{11}^1 = (n-2)$ . The necessary and sufficient condition for a Triangular design to be of the LB type is that either (i)  $r = 2\lambda_1 - \lambda_2$  and  $b = n$  or (ii)  $r = (n-3)\lambda_2 - (n-4)\lambda_1$  and  $b = \frac{1}{2}(n-1)(n-2)$ .

For a Latin square type design with  $i$  constraints,

$$v = n^2, n_1 = i(n-1), n_2 = (n-1)(n-i+1) \text{ and } p_{11}^1 = i(i-3)+n.$$

In order that a Latin Square type design may be of the LB type it is necessary and sufficient that either

$$(i) \quad r = (i-n)(\lambda_1 - \lambda_2) + \lambda_2 \quad \text{and} \quad b = n(n-1) + i$$

$$\text{or} \quad (ii) \quad r = i(\lambda_1 - \lambda_2) + \lambda_2 \quad \text{and} \quad b = i(n-1) + 1,$$

For a Simple design  $\lambda_1 \neq 0$ ,  $\lambda_2 = 0$  Bose and Clatworthy (1955) have shown that the complete class of simple designs with  $\lambda_1 = 1$  and  $k > r \geq 2$  is characterized by the parameters

$$v = k[(r-1)(k-1)+t]/t \quad b = r[(r-1)(k-1)+t]/t$$

$$n_1 = r(k-1), \quad n_2 = (r-1)(k-1)(k-t)/t$$

$$p_{11}^1 = (t-1)(r-1)+(k-2)$$

where

$$1 \leq t \leq r.$$

The necessary and sufficient condition that this is of the LB type is that  $t = r$ .

Using these results, we have picked up all the LB designs amongst the two-associate PBIB designs tabulated by Bose, Clatworthy and Shrikhande (1954) and the designs are presented in § 7.

#### 7. LIST AND PLAN OF ALL LINKED BLOCK DESIGNS WITH $r, k \leq 10$

In this section we give a list of all LB designs (other than the symmetrical BIB designs) requiring at most ten replications and blocks of at most ten plots. The list is arranged in ascending order of the number of varieties, blocks and replications. The number of types of varietal differences estimable with different precisions is denoted by  $m$ , and the number of blocks in which a pair of varieties can occur together is denoted by  $\lambda_1, \lambda_2, \dots, \lambda_m$ . The efficiency factor of the design is denoted by  $E$ . Column 3 gives the nature of the design, and in column 4 a reference is made to the serial number of the design (if it is a two-associate PBIB) in Bose, Clatworthy and Shrikhande (1954) where the plan of the design is given. The plans for designs other than two-associate PBIB are given in the following pages. It is interesting to observe that for all these designs, the efficiency factor is very high, being of the order of 90%.

It also follows easily from the results of Connor (1952) that

$$\frac{2\mu r}{k} - (r + \mu - k) \geq \lambda_j \geq r + \mu - k$$

for

$$j = 1, 2, \dots, m.$$

Hence the number

$$m \leq \frac{2}{k} (k-r)(k-\mu) + 1.$$

This is one of the Irregular Linked Block designs. In this design estimates of differences between pairs of treatments can be divided into four groups, each having a different precision. The peculiarity of this design which can be easily analysed and has an efficiency of 91% but is not partially balanced will be clear from the association matrix given below:



TABLE 6. LIST OF LINKED BLOCK DESIGNS WITH  $r, k \leq 10$

design no.	parameters of the design										nature of the design (13)	reference to plan (14)
	$v$ (2)	$b$ (3)	$r$ (4)	$k$ (5)	$\mu$ (6)	$m$ (7)	$\lambda_1$ (8)	$\lambda_2$ (9)	$\lambda_3$ (10)	$\lambda_4$ (11)	$E$ (12)	
1	6	3	2	4	2	2	2	1	-	-	.88	S 1
2	6	4	2	3	1	2	1	0	-	-	.77	SR 1
3	8	4	3	6	4	2	3	2	-	-	.95	S 7
4	9	3	2	6	3	2	2	1	-	-	.92	S 12
5	10	5	2	4	1	2	1	0	-	-	.79	T 1
6	10	5	3	6	3	2	2	1	-	-	.92	T 15
7	10	5	4	8	6	2	4	3	-	-	.97	S 18
8	10	6	3	5	2	2	2	1	-	-	.91	T 9
9	12	3	2	8	4	2	2	1	-	-	.94	S 22
10	12	4	2	6	2	3	2	1	0	-	.88	** 1
11	12	4	3	9	6	2	3	2	-	-	.97	S 24
12	12	6	5	10	8	2	5	4	-	-	.98	S 28
13	12	9	3	4	1	2	1	0	-	-	.80	SR 20
14	12	9	6	8	5	2	4	3	-	-	.95	SR 26
15	14	7	3	6	2	2	3	1	-	-	.88	S 40
16	14	7	4	8	4	2	4	2	-	-	.94	S 41
17	14	8	4	7	3	2	2	0	-	-	.92	SR 32
18	15	3	2	10	5	2	2	1	-	-	.95	S 46
19	15	6	2	5	1	2	1	0	-	-	.81	T 26



TABLE 6. (Continued)

TABLE 6. (Continued)

design no.	parameters of the design										nature of the design	reference to plan	
	$v$ (1)	$b$ (2)	$r$ (3)	$k$ (4)	$\mu$ (5)	$m$ (6)	$\lambda_1$ (7)	$\lambda_2$ (8)	$\lambda_3$ (9)	$\lambda_4$ (10)			
20	15	6	4	10	6	2	2	3	2	—	(11) (12) (13) (14)	(14)	
21	15	10	4	6	2	2	2	2	1	—	.96	triangular	T 25
22	15	10	6	9	5	2	2	4	3	—	.89	triangular	T 22
23	18	4	2	9	3	3	2	2	1	0	.95	triangular	T 27
24	18	9	4	8	3	4	3	3	2	1	.92	three associate PRIB	** 2
25	18	9	5	10	5	4	4	4	3	2	.91	irregular	*
26	18	10	5	9	4	4	4	4	3	2	.95	irregular	*
27	20	5	2	8	2	3	2	2	1	0	.93	irregular	*
28	20	6	3	10	4	3	3	3	2	1	.89	three associate PRIB	** 3
29	20	16	4	5	1	2	1	1	0	—	.94	three associate PRIB	** 4
30	21	7	2	6	1	2	1	1	0	—	.81	semiregular G.D.	SR 51
31	21	7	3	9	3	2	3	3	1	—	.82	triangular	T 31
32	22	11	5	10	4	2	5	2	2	—	.92	singular G.D.	S 77
33	24	16	6	9	3	3	3	2	0	—	.94	singular G.D.	S 81
34	26	13	3	6	1	2	1	0	—	—	.92	three associate PRIB	*
35	26	13	4	8	2	2	4	1	—	—	.84	simple	S 4
36	28	8	2	7	1	2	1	0	—	—	.90	singular G.D.	S 89
37	30	6	2	10	2	3	2	1	0	—	.84	triangular	T 32
38	30	10	3	9	2	3	2	1	0	—	.90	three associate PRIB	** 5
										.90	irregular	*	

126

# CLASSIFICATION AND ANALYSIS OF LINKED BLOCK DESIGNS

TABLE 6. (Continued)

parameters of the design													
design no.	$v$	$b$	$r$	$k$	$\mu$	$m$	$\lambda_1$	$\lambda_2$	$\lambda_3$	$\lambda_4$	$K$	nature of the design	reference to plan
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
39	30	21	7	10	3	4	3	2	1	0	.93	irregular	*
40	30	25	5	6	1	2	1	0	-	-	.86	semi regular G.D.	SR 70
41	35	15	3	7	1	2	1	0	-	-	.86	simple	S 9
42	36	9	2	8	1	2	1	0	-	-	.85	triangular	T 33
43	36	28	7	9	2	2	2	1	-	-	.91	triangular	T 34
44	42	21	5	10	2	2	5	1	-	-	.91	singular G.D.	S 111
45	45	10	2	9	1	2	1	0	-	-	.86	triangular	T 35
46	50	25	4	8	1	2	1	0	-	-	.88	simple	S 17
47	55	11	2	10	1	2	1	0	-	-	.87	triangular	T 36
48	56	49	7	8	1	2	1	0	-	-	.89	semiregular G.D.	SR 85
49	57	19	3	9	1	2	1	0	-	-	.88	simple	S 18
50	63	28	4	9	1	2	1	0	-	-	.89	simple	S 21
51	70	21	3	10	1	2	1	0	-	-	.89	simple	S 22
52	72	64	8	9	1	2	1	0	-	-	.90	semiregular G. D.	SR 89
53	82	41	5	10	1	2	1	0	-	-	.90	simple	S 25

\*For the plan of designs asterisked see the following pages.

\*\*1 Replace each variety in S1 by two varieties to get the plan for Design no 10

\*\*2 Replace each variety in S1 by three varieties to get the plan for Design no 23

\*\*3 Replace each variety in T1 by two varieties to get the plan for Design no 27.

\*\*4 Replace each variety in T9 by two varieties to get the plan for Design no. 28

\*\*5 Replace each variety in T20 by two varieties to get the plan for Design no. 37

TABLE 7. DESIGN NO. 24 (Irregular)

$r = 18, \quad b = 9, \quad r = 4, \quad k = 8, \quad \mu = 3, \quad E = 0.91,$ $m = 4, \quad \lambda_1 = 3, \quad \lambda_2 = 2, \quad \lambda_3 = 1, \quad \lambda_4 = 0.$								
blocks	plan				plan			
1	1	2	3	4	5	6	7	8
2	1	2	3	9	10	11	12	13
3	1	4	6	9	12	14	16	17
4	1	5	7	10	12	15	16	18
5	2	4	8	10	13	14	16	18
6	2	5	6	11	13	15	16	17
7	3	4	8	11	12	15	17	18
8	3	5	7	9	13	14	17	18
9	6	7	8	9	10	11	14	15

TABLE 8. ASSOCIATION MATRIX OF THE DESIGN NO. 24

		number of blocks in which the pair of treatments (i, j) occur together																
i \ j	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1		2	2	2	2	2	2	1	2	2	1	3	1	1	1	2	1	1
2			2	2	2	2	1	2	1	2	2	1	3	1	1	2	1	1
3				2	2	1	2	2	2	1	2	2	2	1	1	2	1	1
4					1	2	1	3	1	1	1	2	1	2	1	0	2	2
5						2	3	1	1	1	1	1	2	1	2	2	2	2
6							2	2	2	1	1	1	2	1	2	2	2	2
7								2	2	2	1	2	1	1	2	2	2	0
8									2	2	1	1	1	2	2	1	1	2
9										1	2	2	1	1	2	2	1	2
10											2	2	2	2	3	1	1	2
11												2	2	2	2	2	0	2
12													2	2	1	3	1	2
13														1	1	2	2	2
14															2	1	2	2
15																1	2	2
16																	2	2
17																		2
18																		

# CLASSIFICATION AND ANALYSIS OF LINKED BLOCK DESIGNS

TABLE 9. DESIGN NO. 25 (Irregular)

$v = 18, \quad b = 9, \quad r = 5 \quad k = 10 \quad \mu = 5 \quad E = 0.95$ $m = 4 \quad \lambda_1 = 4 \quad \lambda_2 = 3 \quad \lambda_3 = 2 \quad \lambda_4 = 1$										
plan										
blocks	1	7	8	9	10	13	14	15	17	18
1	1	7	8	9	10	13	14	15	17	18
2	2	3	6	8	9	10	12	14	16	17
3	1	2	5	7	8	9	11	12	13	16
4	2	3	5	6	7	10	11	13	14	15
5	3	4	5	9	10	12	13	15	16	18
6	1	2	4	5	6	8	12	14	15	18
7	1	3	4	7	11	12	14	15	16	17
8	1	2	3	4	6	9	11	13	17	18
9	4	5	6	7	8	10	11	16	17	18

TABLE 10. DESIGN NO. 26 (Irregular)

$v = 18 \quad b = 10 \quad r = 5 \quad k = 9 \quad \mu = 4 \quad E = 0.93$ $m = 4 \quad \lambda_1 = 4 \quad \lambda_2 = 3 \quad \lambda_3 = 2 \quad \lambda_4 = 1$									
plan									
blocks	1	2	3	4	5	6	7	8	9
1	1	2	3	4	5	6	7	8	9
2	1	2	3	4	10	11	12	13	14
3	1	2	5	6	10	11	15	16	17
4	1	3	7	8	10	12	15	16	18
5	1	4	7	9	11	13	16	17	18
6	2	3	5	7	13	14	15	17	18
7	2	4	6	9	12	14	15	16	18
8	4	5	6	8	10	12	13	17	18
9	3	5	8	9	11	12	14	16	17
10	6	7	8	9	10	11	13	14	15



TABLE 11. DESIGN NO 33  
(Partially Balanced with Three Associate-Classes)

$v = 24, b = 16, r = 6, k = 9, \mu = 3, E = 0.92$									
$m = 3, \lambda_1 = 3, \lambda_2 = 2, \lambda_3 = 0, n_1 = 4, n_2 = 18, n_3 = 1$									
blocks	plan								
1	1	2	3	4	5	6	7	8	9
2	1	2	3	10	11	12	13	14	15
3	4	5	6	10	11	12	16	17	18
4	7	8	9	13	14	15	16	17	18
5	1	4	7	10	13	16	19	20	21
6	1	5	8	11	14	16	19	22	23
7	2	4	8	11	13	17	20	22	24
8	2	5	7	10	14	17	21	23	24
9	3	6	9	11	14	17	19	20	21
10	3	4	7	12	15	17	19	22	23
11	1	6	7	12	14	18	20	23	24
12	1	4	9	11	15	18	21	23	24
13	2	5	8	12	15	18	19	20	21
14	2	6	9	10	13	18	19	22	23
15	3	5	9	10	15	16	20	22	24
16	3	6	8	12	13	16	21	23	24

This is a partially balanced design with three-associate classes. The values of the  $p_{js}^i$  parameters are given below:

$$\begin{bmatrix} p_{11}^1 = 2 & p_{12}^1 = 0 & p_{13}^1 = 1 \\ & p_{22}^1 = 18 & p_{23}^1 = 0 \\ & & p_{33}^1 = 0 \end{bmatrix}$$

$$\begin{bmatrix} p_{11}^2 = 0 & p_{12}^2 = 4 & p_{13}^2 = 0 \\ & p_{22}^2 = 12 & p_{23}^2 = 1 \\ & & p_{33}^2 = 0 \end{bmatrix}$$

$$\begin{bmatrix} p_{11}^3 = 4 & p_{12}^3 = 0 & p_{13}^3 = 0 \\ & p_{22}^3 = 18 & p_{23}^3 = 0 \\ & & p_{33}^3 = 0 \end{bmatrix}$$

# CLASSIFICATION AND ANALYSIS OF LINKED BLOCK DESIGNS

TABLE 12. DESIGN NO. 38 (Irregular)

$v = 30, b = 10, r = 3, k = 9, \mu = 2, E = 0.90$ $m = 3, \lambda_1 = 2, \lambda_2 = 1, \lambda_3 = 0.$									
blocks	plan								
	1	2	3	4	5	6	7	8	9
1	1	2	10	11	12	13	14	15	16
2	1	3	10	17	18	19	20	21	22
3	2	4	11	17	18	23	24	25	26
4	3	5	12	13	19	23	24	27	28
5	4	6	10	14	19	25	27	29	30
6	5	7	14	15	17	20	26	28	29
7	6	8	12	16	18	21	26	28	30
8	7	9	13	15	21	22	23	25	30
9	8	9	11	16	20	22	24	27	29
10									

TABLE 13. DESIGN NO. 39 (Irregular)

$v = 30, b = 21, r = 7, k = 10, \mu = 3, E = 0.93,$ $m = 4, \lambda_1 = 3, \lambda_2 = 2, \lambda_3 = 0.$										
blocks	plan									
	1	2	4	8	9	11	15	16	18	28
1	1	2	4	8	9	10	12	16	17	22
2	2	3	5	9	10	12	16	17	19	27
3	1	3	7	8	10	14	15	17	21	26
4	2	6	7	9	13	14	16	20	21	26
5	1	5	6	8	12	13	15	19	20	25
6	4	5	7	11	12	14	18	19	21	24
7	3	4	6	10	11	13	17	18	20	23
8	4	5	9	14	17	20	25	27	28	30
9	3	4	8	13	16	19	24	26	27	30
10	2	3	12	14	15	18	23	25	26	30
11	1	2	11	13	17	21	22	24	25	30
12	1	7	10	12	16	20	23	24	28	30
13	6	7	9	11	15	19	22	23	27	30
14	5	6	8	10	18	21	22	26	28	30
15	3	6	11	12	16	21	25	27	28	29
16	2	5	10	11	15	20	24	26	27	29
17	1	4	9	10	19	21	23	25	26	29
18	3	7	8	9	18	20	22	24	25	29
19	2	6	8	14	17	19	23	24	28	29
20	1	5	13	14	16	18	22	23	27	29
21	4	7	12	13	15	17	22	26	28	29

## REFERENCES

- BOSE, R. C. AND NAIR, K. R. (1939): Partially balanced incomplete block designs. *Sankhyā* 4, 337-372.
- AND CONNOR, W. S. (1952): Combinatorial properties of group divisible incomplete block designs. *Ann. Math. Stat.*, 23, 367-383.
- BOSE, R. C. AND SHIMAMOTO, T. (1952): Classification and analysis of partially balanced incomplete block designs with two associate classes. *J. Amer. Stat. Ass.*, 47, 151-184.
- BOSE, R. C. CLATWORTHY, W. H. AND SHRIKHANDE, S. S. (1954): Tables of partially balanced designs with two associate classes. *Tech. Bull.*, No. 107, Reprint Series No. 50, University of North Carolina.
- BOSE, R. C. AND CLATWORTHY, W. H. (1955): Some classes of partially balanced designs. *Ann. Math. Stat.*, 26, 212-232.
- CONNOR, W. S. (1952): On the structure of balanced incomplete block designs. *Ann. Math. Stat.*, 23, 57-71.
- AND CLATWORTHY, W. H. (1954): Some theorems for partially balanced designs. *Ann. Math. Stat.*, 25, 100-112.
- ROY, P. M. (1954): On the method of inversion in the construction of partially balanced incomplete block designs from the corresponding B.I.B. designs. *Sankhyā* 14, 39-52.
- SHRIKHANDE, S. S. (1952): On the dual of some balanced incomplete block designs. *Biometrics*, 8, 62-72.
- YODEN, W. G. (1951): Linked blocks : a new class of incomplete block designs (Abstract). *Biometrics*, 7, 124.

# ON THE DUAL OF A PBIB DESIGN, AND A NEW CLASS OF DESIGNS WITH TWO REPLICATIONS

By C. S. RAMAKRISHNAN  
*Indian Statistical Institute, Calcutta*

## 1. INTRODUCTION AND SUMMARY

By interchanging blocks and varieties in a given class of designs we get a new class of designs called the dual of the original class. Bose and Nair (1939) gave examples of PBIB designs obtained by dualizing some BIB designs. Youden (1951) investigated these further. He called the dual of a BIB design by the suggestive name of Linked Block designs. Elsewhere in this issue of *Sankhyā* Roy and Laha (1956) have made exhaustive study, classification and enumeration of all Linked Block designs with  $r, k \leq 10$ . In this paper, we give the analysis and structure of the dual of a PBIB design with two associate classes. By dualizing a simple class of designs with 2 plots per block, we also derive a new class of useful designs with two replications. These turn out to be PBIB designs with five associate classes, but the analysis of these designs by the dual method turns out to be extremely simple, whereas a straightforward analysis of these as PBIB designs with five associate classes would be tedious.

## 2. GENERAL REMARKS ON THE DUAL METHOD

Dualization of known types of designs sometimes leads us to new designs and sometimes only yields designs already known. For example, most of the Linked Block designs given by Roy and Laha (1956) turn out to be PBIB designs with two or three associate classes. But even in these cases, the dual method gives us a simpler way of analyzing these designs. Before starting the analysis of a design, it is always worthwhile to see whether the dual is easier to be analyzed, in which case we may profitably follow the dual method of analysis. The labour involved in preparing the analysis of variance table for the dual of a design is almost the same as that involved in the analysis of the original design. But there is additional labour in getting the standard errors of treatment contrasts. We have to impose restrictions on the original design to start with, so that in the dual design we may have only a small number of types of varietal differences with different precisions, and neat expressions for the standard errors.



## 3. THE DUAL OF A TWO ASSOCIATE PBIB DESIGN

Consider a PBIB design with two associate classes. We adopt the standard notation (except for an asterisk) of Bose, Clatworthy and Shrikhande (1954). Let the parameters of the design be  $v^*, r^*, k^*, b^*, \lambda_1^*, \lambda_2^*, n_1^*, n_2^*$ . Let us also introduce the constants  $c_1$  and  $c_2$  defined by them. The normal equations for treatment effects  $\{t_i^*\}$  take the form (again adopting the notation of the above authors)

$$r^* (k^* - 1) t_i^* = (k^* - c_2) Q_i^* + (c_1 - c_2) S_1(Q_i^*)$$

where

$Q_i^*$  = adjusted yield of the  $i$ -th treatment.

$S_1(Q_i^*)$  = sum of the  $Q_i^*$ 's of all the first associates of the  $i$ -th treatment.

The dual of this design will have parameters

$$v = b^*, \quad r = k^*, \quad k = r^*, \quad b = v^*.$$

Let  $[n_{ij}]_{b \times v}$  be the incidence matrix of this design,

$T_j$  = total yield of the  $j$ -th treatment,

$Q_j$  = the corresponding adjusted yield.

$B_i$  = total yield of the  $i$ -th block,

$P_i$  = the corresponding adjusted block total.

Then the normal equations for treatment effects  $\{t_j\}$  and block effects  $\{b_i\}$  take one of the two following forms (3.1) or (3.2):

$$\left. \begin{aligned} T_j &= r t_j + \sum_i n_{ij} b_i \\ B_i &= k b_i + \sum_j n_{ij} t_j \end{aligned} \right\} \quad \dots (3.1)$$

$$\left. \begin{aligned} Q_j &= T_j - \frac{1}{k} \sum_i n_{ij} B_i = r(k-1) t_j - \frac{1}{k} \sum_{\substack{j'=1 \\ j' \neq j}}^{j'=v} \lambda_{jj'} t_{j'} \\ P_i &= B_i - \frac{1}{r} \sum_j n_{ij} T_j = k(r-1) b_i - \frac{1}{r} \sum_{i' \neq i} \mu_{ii'} b_{i'} \end{aligned} \right\} \quad \dots (3.2)$$

where  $\lambda_{jj'}$  = number of blocks in which treatments  $j$  and  $j'$  occur together, and  $\mu_{ii'}$  = number of varieties common to blocks  $i$  and  $i'$ .

If the given design is the dual of the original PBIB design, the equations for block effects take the form

$$k(r-1) b_i = (r - c_2) P_i + (c_1 - c_2) S_1(P_i). \quad \dots (3.3)$$

We compute  $\{b_i\}$  by the above formula.

# A NEW CLASS OF DUALS OF PBIB DESIGNS

The over-all analysis of variance table may now be set up.

TABLE 1. ANALYSIS OF VARIANCE

source	formula	s.s.	d.f.	s.s.	formula	source
blocks (ignoring treatments)	$\frac{1}{k} \sum B_i^2 - \text{c.f.}$	$B$	$b-1$	$BE$	$\sum b_i P_i$	blocks (eliminating treatments)
treatments (eliminating blocks)	$T - E - B$	$VE$	$v-1$	$V$	$\frac{1}{r} \sum T_j^2 - \text{c.f.}$	treatments (ignoring blocks)
error		$E$	$bk-b-v+1$	$E$	$T - V - BE$	error
total		$T$	$bk-1$	$T$		total

To obtain estimates of treatment contrasts, we compute  $t_j$  from (3.1) which may be rewritten as

$$t_j = \frac{1}{r} [T_j - \sum_i n_{ij} b_i]. \quad \dots (3.4)$$

A check on the calculations is got by computing s.s. due to treatments (eliminating blocks) by the alternative formula  $\sum_j t_j Q_j$ . Estimate of any linear contrast of treatment effects is the corresponding linear function of the  $t_j$ 's.

We now turn to the problem of getting the standard error of any contrast. The method usually given in books is to add one more equation to (3.2) say  $\sum t_j = 0$  and invert the matrix of coefficients. It is not recognised that these can be got in a simple and elegant way through the "Q" technique of Rao (1952). Rao's method is this. Obtain any solution of the normal equations, expressing  $t_j$  as a linear function of  $T_j$ 's and  $B_i$ 's. Then if  $c_{jj'}$  be the coefficient of  $T_{j'}$  in this expression, then the variance of any linear function of the  $t_j$ 's is given by

$$V(\sum l_j t_j) = (\sum_j \sum_{j'} l_j l_{j'} c_{jj'}) \sigma^2 \quad \dots (3.5)$$

where  $\sigma^2$  is the intra block error variance. Thus

$$V(t_j - t_{j'}) = (c_{jj} + c_{j'j'} - 2c_{jj'}) \sigma^2.$$

Hence the number of distinct types of standard errors will depend on the number of distinct  $c_{jj'}$ .

In our case, following Rao's method

$$\begin{aligned} t_j &= \frac{1}{r} \left\{ T_j - \sum_i n_{ij} b_i \right\} \\ &= \frac{1}{r} \left[ T_j - \frac{r-c_2}{k(r-1)} \left\{ \sum_i n_{ij} \left( B_i - \frac{1}{r} \sum_{j'} n_{ij'} t_{j'} \right) \right\} \right] + \\ &\quad + \frac{c_1 - c_2}{k(r-1)} \left[ \sum_i n_{ij} S_1 \left( B_i - \frac{1}{r} \sum_{j'} n_{ij'} t_{j'} \right) \right]. \end{aligned}$$

Then 
$$c_{jj} = \text{coefficient of } T_j = \frac{1}{r} \left\{ 1 + \frac{r-c_2}{k(r-1)} + \frac{c_1-c_2}{k(r-1)} v_{jj} \right\}$$

and 
$$c_{jj'} = \text{coefficient of } T_{j'} = \frac{1}{r^2} \left\{ \frac{r-c_2}{k(r-1)} \lambda_{jj'} + \frac{c_1-c_2}{k(r-1)} v_{jj'} \right\}$$

where  $v_{jj'}$  = number of times  $j'$  occurs in the first associates of all the blocks in which  $j$  occurs. Thus  $c_{jj'}$  depends on  $\lambda_{jj'}$ ,  $v_{jj}$  and  $v_{jj'}$ .

For a general PBIB design, calculation of  $v_{jj'}$  is tedious, and we have to impose restrictions to limit the number of distinct types of errors. A very effective restriction is imposed by taking  $r = 2$ .

#### 4. A NEW CLASS OF DESIGNS

Here we derive a new class of designs with  $r = 2$  by dualizing the following simple class of designs. Take  $v^*$  to be even  $= 2m$ , write down all pairs and omit the pairs of the form  $(2i-1, 2i)$

$$\begin{array}{ll} (13) & (14) \dots\dots (1 \quad 2m) \\ (23) & (24) \dots\dots (2 \quad 2m) \\ & (35) \dots\dots (3 \quad 2m) \\ & \dots\dots\dots \\ & (2m-2 \quad 2m-1) \quad (2m-2 \quad 2m). \end{array}$$

This is a Group Divisible design (Bose et al, 1954)

$$\begin{array}{ll} \text{groups} & v^* = mn; \quad m = m; n = 2; \\ 1 \quad 2 & \lambda_1^* = 0; \quad \lambda_2^* = 1 \\ 3 \quad 4 & n_1 = 1; \quad n_2 = 2(m-1) \\ \dots \quad \dots & \\ 2m-1 \quad 2m & k^* = 2; \quad r^* = v-2; \quad b^* = 2m(m-1). \end{array}$$

Let  $i_a$  be the associate of the  $i$ -th treatment

$$\begin{aligned} i_a &= i+1 && \text{if } i \text{ is odd} \\ &= i-1 && \text{if } i \text{ is even.} \end{aligned}$$

The dual of this design has  $v = 2m(m-1)$ ;  $k = 2(m-1)$ ;  $r = 2$ ;  $b = 2m$ .

Following the method of § 2, normal equations for block effects  $\{b_i\}$  and treatment effects  $\{t_j\}$  are

$$\left. \begin{aligned} b_i &= \frac{2(k+1)}{k(k+2)} P_i - \frac{2}{k(k+2)} P_{i_a} \\ 2t_j &= T_j - b_j^{(p)} - b_j^{(q)} \end{aligned} \right\} \dots (4.1)$$

where  $b_j^{(p)}$ ,  $b_j^{(q)}$  are the effects of the two blocks in which the  $j$ -th treatment occurs.

## A NEW CLASS OF DUALS OF PBIB DESIGNS

Solving for  $b_i$  and then for  $t_j$  we can get estimates of any treatment contrasts, and s.s. due to treatments. The analysis of variance table is easily set up, as in § 2.

To get the standard errors, we compute  $c_{jj}$  and  $c_{jj'}$ . We get

$$\left. \begin{aligned} v_{jj} &= 0; \quad c_{jj} = \frac{1}{2} + \frac{k+1}{k(k+2)} = c \text{ (say)} \\ c_{jj'} &= \frac{k+1}{2(k+2)k} \lambda_{jj'} - \frac{1}{2(k+2)k} v_{jj'} \\ V(t_j - t_{j'}) &= 2(c - c_{jj'})\sigma^2. \end{aligned} \right\} \dots (4.2)$$

There are five distinct types of errors for treatment differences corresponding to the following five combinations of values of  $\lambda_{jj'}$  and  $v_{jj'}$ .

$$\left. \begin{array}{ll} (1) & \lambda_{jj'} = 1; \quad v_{jj'} = 0 \\ (2) & \lambda_{jj'} = 1; \quad v_{jj'} = 1 \\ (3) & \lambda_{jj'} = 0; \quad v_{jj'} = 2 \\ (4) & \lambda_{jj'} = 0; \quad v_{jj'} = 1 \\ (5) & \lambda_{jj'} = 0; \quad v_{jj'} = 0 \end{array} \right\} \dots (4.3)$$

Actually, the new design turns out to be a PBIB design with five associate classes, if we define association between two treatments  $j$  and  $j'$  by the above five conditions on  $\lambda_{jj'}$  and  $v_{jj'}$ . We shall say that  $j$  and  $j'$  are first associates if  $\lambda_{jj'} = 1, v_{jj'} = 0$ ; second associates if  $\lambda_{jj'} = 1, v_{jj'} = 1$ ; and so on. An easy way of writing down the association scheme is as follows.

Let  $p$  and  $q$  be the blocks in which  $j$  occurs and  $p_a, q_a$  the respective associates of these blocks. Blocks form a Group Divisible design with  $\lambda_1^* = 0, \lambda_2^* = 1$  with groups (12), (34), ...,  $(2m-1, 2m)$ , and block size  $k = 2(m-1)$ . Consider the four blocks  $p, p_a, q, q_a$ . We shall give a method of writing down all the associates of treatment  $j$  by a mere inspection of these four blocks. The blocks  $p$  and  $q$  have one variety in common namely  $j$ . Let  $j_{pq_a}$  and  $j_{p_aq}$  be the varieties common to  $p, q_a$  and  $p_a, q$  respectively. These two varieties are those for which  $\lambda_{jj'} = 1, v_{jj'} = 1$ . Thus for any treatment  $j$ , there are exactly two second associates  $\lambda_2 = 1; n_2 = 2$ . The remaining  $2(k-2)$  varieties of  $p$  and  $q$  (omitting the above three namely  $j, j_{pq_a}$  and  $j_{p_aq}$ ) are those for which  $\lambda_{jj'} = 1, v_{jj'} = 0$ . These are first associates,  $\lambda_1 = 1, n_1 = 2(k-2)$ . Blocks  $(p_a, q_a)$  have exactly one variety in common and this cannot occur in either  $p$  or  $q$ . Call this  $j_{p_aq_a}$ . This satisfies  $\lambda_{jj'} = 0, v_{jj'} = 2$ . Thus



there is exactly one third associate:  $\lambda_3 = 0, n_3 = 1$ . The remaining  $2(k-2)$  varieties in the blocks  $p_a$  and  $q_a$  are the fourth associates  $\lambda_{ij} = 0, n_{ij} = 1$ ;  $\lambda_4 = 0, n_4 = 2(k-2)$ . The varieties not occurring in any of the above four blocks are the fifth associates:  $\lambda_5 = 0, n_5 = 2(m-2)(m-3)$ .

Following the above method, the association scheme and the parameters  $[p_{ij}^k]$  can be written down. For  $m = 3$  we have  $n_3 = 0$ . There are only four associate classes and we get a design given by Nair (1951). For  $m=4, 5$ , etc., we get new designs with five associate classes. For  $m=6, k=10$  and higher values of  $m$  are not desirable.

It may be noted that the analysis of these designs as PBIB with five associate classes is tedious, whereas the dual method of analysis is extremely simple, illustrating the usefulness of the dual method of analysing designs.

### 5. NUMERICAL EXAMPLE

Here we give numerical details of analysis of one of our new designs, design No. 2, in our list. Parameters of this design are  $r = 24, m = 4; b = 8; k = 6$ .  $\lambda_1 = \lambda_2 = 1, \lambda_3 = \lambda_4 = \lambda_5 = 0, n_1 = n_4 = 8, n_2 = 2, n_3 = 1, n_5 = 4$ .

TABLE 2. PLAN AND YIELDS

block no.	varieties and yields*											
	(1)	1.5	(2)	3.4	(3)	3.5	(4)	7.0	(5)	6.8	(6)	7.2
1	(1)	1.5	(2)	3.4	(3)	3.5	(4)	7.0	(5)	6.8	(6)	7.2
2	(7)	2.8	(8)	5.7	(9)	4.0	(10)	4.9	(11)	7.6	(12)	8.1
3	(1)	3.4	(7)	4.7	(13)	4.4	(14)	4.4	(15)	5.6	(16)	6.1
4	(2)	4.4	(8)	6.1	(17)	7.8	(18)	10.3	(19)	5.3	(20)	6.9
5	(3)	7.4	(9)	9.4	(13)	4.9	(17)	11.3	(21)	8.5	(22)	9.3
6	(4)	10.2	(10)	9.7	(14)	8.0	(18)	11.0	(23)	10.5	(24)	12.3
7	(5)	11.4	(11)	9.6	(15)	7.6	(19)	6.1	(21)	9.1	(23)	10.3
8	(6)	10.2	(12)	11.8	(16)	8.0	(20)	7.7	(22)	7.3	(24)	11.1

\*Varieties are indicated within brackets. Yields are written by the side.

TABLE 3. ESTIMATION OF BLOCK EFFECTS

blocks	I	II	III	IV	V	VI	VII	VIII	sum for checks
$B_i$ (unadjusted block totals)	29.4	33.1	28.6	40.8	50.8	61.7	54.1	56.1	354.6
$\sum n_{ij}T_j$ (sum of totals of treatments in that block)	76.4	84.4	61.4	86.0	86.9	109.7	98.4	106.0	709.2
$P_i$ (adjusted block totals)	-8.8	-9.1	-2.1	-2.2	7.35	6.85	4.9	3.1	0
$b_i$ (estimate of block effects)	-2.2	-2.3	-0.5	-0.6	1.9	1.7	1.3	0.7	0
Computational checks: $\sum B_i = G = \text{total yield}, \sum \sum n_{ij}T_j = 2G, \sum P_i = 0 = \sum b_i$									

# A NEW CLASS OF DUALS OF PBIB DESIGNS

TABLE 4. ANALYSIS OF VARIANCE

source	s.s.	m.s.	F	d.f.	F	m.s.	s.s.	source
blocks (ignoring treatments)	197.74			7	*16.7	10.88	76.19	blocks (eliminating varieties)
varieties (eliminating blocks)	†135.84	5.91	*9.06	23			257.30	varieties (ignoring blocks)
error	11.09	0.652		17		0.652	†11.09	error
total	344.67			47			344.67	total

† obtained by subtraction. \* significant at 1%.

TABLE 5. ESTIMATION OF TREATMENT EFFECTS

$j$ (treatment no.)		1	2	3	4	5	6	7	8	9	10	11	12
$T_j$	(unadjusted treatment totals)	4.9	7.8	10.9	17.2	18.2	17.4	7.5	11.8	13.4	14.6	17.2	19.9
$\Sigma n_{ij}b_i$	(sum of the block effects in which the treatment occurs)	-2.7	-2.7	-0.3	-0.5	-0.9	-1.5	-2.8	-2.8	-0.4	-0.6	-1.0	-1.6
$t_j$	(estimate of treatment effects)	3.8	5.3	5.6	8.8	9.6	9.4	5.2	7.3	6.9	7.6	9.1	10.8

$j$ (treatment no.) (contd.)		13	14	15	16	17	18	19	20	21	22	23	24
$T_j$	(unadjusted treatment totals)	9.3	12.4	13.2	14.1	19.1	21.3	11.4	14.6	17.6	16.6	20.8	23.4
$\Sigma n_{ij}b_i$	(sum of the block effects in which the treatment occurs)	1.3	1.2	0.8	0.2	1.3	1.1	0.8	0.2	3.2	2.6	3.0	2.4
$t_j$	(estimate of treatment effects)	4.0	5.6	6.2	7.0	8.9	10.1	5.3	7.2	7.2	7.0	8.9	10.5

Computational checks:  $\Sigma T_j = G$ ;  $\Sigma \Sigma n_{ij}b_i = 0$ ;  $2\Sigma t_j = G$ .

TABLE 6. STANDARD ERROR OF VARIETAL DIFFERENCES

type of associates	$\lambda_{jj'}$	$r_{jj'}$	estimate of $V(t_j - t_{j'})$	critical dif- ference at 5% level of significance
1	1	0	0.75	1.83
2	1	1	0.76	1.84
3	0	2	0.87	1.97
4	0	1	0.85	1.95
5	0	0	0.84	1.93

The above table can be used to test the significance of any varietal difference, first by determining the type of associates, and then comparing with the corresponding critical difference. For example  $t_7 - t_1 = 1.4$ . This does not exceed 1.84. the critical difference for second associates, and hence is not significant.

#### 6. PLANS OF NEW DESIGNS

Here we give plans and parameters of all designs of the new class, with  $k \leq 10$ . There are four such corresponding to  $m = 3, 4, 5, 6$ . We also give a table giving constants for calculating standard errors.

We tabulate  $K_{jj'}$ , where

$$V(t_j - t_{j'}) = K_{jj'} \sigma^2,$$

for the five types of associates.

TABLE 7. VALUES OF PARAMETERS

$r = 2; \lambda_1 = \lambda_2 = 1; \lambda_3 = \lambda_4 = \lambda_5 = 0; b = 2m; k = 2(m-1);$ $v = 2m(m-1); n_1 = n_4 = 4(m-2); n_5 = 2(m-2)(m-3)$						
design no.	$m$	$b$	$k$	$v$	$n_1 = n_4$	$n_5$
1	3	6	4	12	4	0
2	4	8	6	24	8	4
3	5	10	8	40	12	12
4	6	12	10	60	16	24

TABLE 8. PLANS OF FOUR NEW DESIGNS

*Numbers of blocks are indicated within brackets. Treatments in the block are written by the side.*

DESIGN NO. 1.									
(1)	1	2	3	4	(2)	5	6	7	8
(3)	1	5	9	10	(4)	2	6	11	12
(5)	3	7	9	11	(6)	4	8	10	12

# A NEW CLASS OF DUALS OF PBIB DESIGNS

TABLE 8. (Continued)

TABLE 3. (Contd.)

DESIGN NO. 2													
(1)	1	2	3	4	5	6	(2)	7	8	9	10	11	12
(3)	1	7	13	14	15	16	(4)	2	8	17	18	19	20
(5)	3	9	13	17	21	22	(6)	4	10	14	18	23	24
(7)	5	11	15	19	21	23	(8)	6	12	16	20	22	24

DESIGN NO. 3																	
(1)	1	2	3	4	5	6	7	8	(2)	9	10	11	12	13	14	15	16
(3)	1	9	17	18	19	20	21	22	(4)	2	10	23	24	25	26	27	28
(5)	3	11	17	23	29	30	31	32	(6)	4	12	18	24	33	34	35	36
(7)	5	13	19	25	29	33	37	38	(8)	6	14	20	26	30	34	39	40
(9)	7	15	21	27	31	35	37	39	(10)	8	16	22	28	32	36	38	40

DESIGN NO. 4																					
(1)	1	2	3	4	5	6	7	8	9	10	(2)	11	12	13	14	15	16	17	18	19	20
(3)	1	11	21	22	23	24	25	26	27	28	(4)	2	12	29	30	32	32	33	34	35	36
(5)	3	13	21	29	37	38	39	40	41	42	(6)	4	14	22	30	43	44	45	46	47	48
(7)	5	15	23	31	37	43	49	50	51	52	(8)	6	16	24	32	38	44	53	54	55	56
(9)	7	17	25	33	39	45	49	53	57	58	(10)	8	18	26	34	40	46	50	54	59	60
(11)	9	19	27	35	41	47	51	55	57	59	(12)	10	20	28	36	42	48	52	56	58	60

We give below  $K_{jj'}$  where  $V(t_j - t_{j'}) = K_{jj'} \sigma^2$  for the five types of associates. For a randomised block experiment with  $r = 2$ ,  $V(t_j - t_{j'}) = \sigma^2$ . Hence  $1/K_{jj'}$  stands for the "efficiency factor" as usually understood.

TABLE 9. CONSTANTS  $K_{jj'}$  FOR THE CALCULATION OF STANDARD ERROR

type of associates	design number			
	1	2	3	4
(1)	(2)	(3)	(4)	(5)
1	1.21	1.15	1.11	1.09
2	1.25	1.17	1.12	1.10
3	1.50	1.33	1.25	1.20
4	1.46	1.31	1.24	1.19
5	1.42	1.29	1.22	1.18



## CONCLUSION

My sincere thanks are due to Dr. C. R. Rao for suggesting the problem and for guidance and to Shri J. Roy and Shri R. G. Laha for helpful discussions.

## REFERENCES

- BOSE, R. C., CLATWORTHY, W. H. and SHRIKHANDE, S. S. (1954): *Tables of Partially Balanced Designs with Two Associate Classes*. University of North Carolina.
- BOSE, R. C. AND NAIR, K. R. (1939): Partially balanced incomplete block designs. *Sankhyā*, **4**, 337
- NAIR, K. R. (1951): Some two-replicate pbib designs. *Cal. Stat. Ass. Bull.*, **3**, 174-176.
- RAO, C. R. (1952): *Advanced Statistical Methods in Biometric Research*. John Wiley & Sons, New York.
- ROY, J. and LAHA, R. G. (1956): Classification and analysis of linked block designs. *Sankhyā*, **17**, 115-132.
- YODEN, W. J. (1951): Linked blocks: a new class of incomplete block designs. (Abstract), *Biometrics*, **7**, 124.

*Paper received : February, 1956.*

# FRACTIONAL REPLICATION IN ASYMMETRICAL FACTORIAL DESIGNS AND PARTIALLY BALANCED ARRAYS

By I. M. CHAKRAVARTI

*Indian Statistical Institute, Calcutta*

## 1. INTRODUCTION

Hypercubes of strength " $d$ " were defined by Rao (1946). Later Rao (1947) extended the definition of hypercubes of strength  $d$  to cover a wider class of arrays called orthogonal arrays. Rao (1946) has shown how these hypercubes may be used in the construction of a system of confounded designs which accommodates maximum number of factors and preserves main effects and interactions up to the order  $(d-1)$  provided a hypercube of strength " $d$ " exists in the case of a symmetrical factorial experiment. It has also been shown there that hypercubes of strength 2 supply balanced confounded designs for asymmetrical factorial experiments defined by Nair and Rao (1941, 1942a, 1942b) and later (Nair and Rao, 1948) treated in detail by the same authors.

Plackett and Burman (1946) constructed a class of designs called multifactorial designs which accommodate maximum number of factors and preserve only the main effects.

Rao (1947) has shown how orthogonal arrays of strength  $d$  may be made to yield multifactorial designs which will allow estimation of main effects and interactions of order upto  $k$  ( $d > k$ ) when higher order interactions are absent. He has also used orthogonal arrays in the construction of block designs for symmetrical factorial experiments involving only a sub-set of treatment combinations and preserving main effects and interaction up to a given order, when higher order interactions are absent.

The existence of block designs allows construction of fractional replication in the case of symmetrical factorial experiments. The method of actual construction of fractional replicates using orthogonal arrays has been treated fully by Rao (1950).

In this paper, the problem of construction of arrangements of fractional replication in asymmetrical factorial designs has been considered. It has been shown that orthogonal arrays may be used to obtain fractional replications in some of the important asymmetrical factorial experiments which find ready application in actual fields of research like industrial experimentation. These fractional replicate designs lead to a considerable saving in the number of experiments to be conducted or observations to be made. Method of construction of these designs are flexible to a certain extent to suit the needs of the varying nature of experimental enquiries. Experimental situations which have actually occurred in practice in the fields of industrial experimentation are considered. A list of useful designs has been supplied.

A new class of arrays called partially balanced arrays has been defined. The combinatorial problem and analysis of designs derived from these partially balanced arrays are given. These designs economise considerably the amount of experimental material to be used in the experiment. These will be found useful in those situations where the most economic design does not exist.

## 2. ORTHOGONAL ARRAYS AND FRACTIONAL REPLICATION

Let  $A$  be a matrix with  $m$  rows and  $N$  columns, elements of the matrix being the integers  $0, 1, 2, \dots, s-1$ . If amongst the  $N$  columns of any of  $\binom{m}{d}$   $d$ -row submatrices from  $A$ , all  $s^d$   $d$ -tuples occur equal number of times, say  $\lambda$  times, then  $A$  is an orthogonal array with  $N$  assemblies,  $m$  constraints, strength  $d$  and index  $\lambda$ , symbolically denoted by  $(N, m, s, d, \lambda)$ . Then it follows that  $N = \lambda s^d$ . If  $N = s^d$  then  $\lambda = s^{d-d}$  and such an array is called a hypercube. The following two general inequalities due to Rao (1947) connect the parameters

$$N-1 \geq \binom{m}{1}(s-1) + \binom{m}{2}(s-1)^2 + \dots + \binom{m}{\frac{1}{2}d}(s-1)^{\frac{1}{2}d}$$

when  $d$  is even and

$$N-1 \geq \binom{m}{1}(s-1) + \dots + \binom{m}{\frac{1}{2}(d-1)}(s-1)^{\frac{1}{2}(d-1)} + \binom{m-1}{\frac{1}{2}(d-1)}(s-1)^{\frac{1}{2}(d+1)}$$

when  $d$  is odd.

For a symmetrical factorial experiment involving  $m$  factors each at  $s$  levels,  $N$  columns of an array are identifiable with  $N$  treatment combinations or assemblies,  $m$  rows stand for  $m$  factors and an entry stands for the level of a factor against which it is shown. These  $N$  assemblies then form a sub-set of  $s^m$  possible assemblies of the complete factorial experiment. From a complete factorial experiment all main effects and interactions of all orders upto  $(m-1)$  are estimable but these take up all the  $s^m - 1$  degrees of freedom leaving none for error. In such situations, one may use estimates of error variance from previous experiences or one may derive a valid estimate of error variance assuming certain higher order interactions to be absent. Sometimes, it is not possible to set up even a single complete replication of a factorial experiment. To get over this difficulty Finney (1945) introduced fractionally replicated designs which using only a sub-set (properly chosen) of  $s^m$  assemblies provide estimates of main effects and lower order interactions on the assumption that higher order interactions are absent.

Rao (1947) has proved that a sub-set of  $N$  assemblies forming an orthogonal array  $(N, m, s, d+k-1, \lambda)$  yields a fractionally replicated design from which :

(i) all the main effects and interactions upto order  $(k-1)$  can be measured when interactions of order equal to and greater than  $d-1$  are absent,



## FRACTIONAL REPLICATES AND PARTIALLY BALANCED ARRAYS

(ii) the expressions for main effects and interactions are simply obtained from the usual definitions by retaining only the treatment combinations present in the array and these expressions belonging to different contrasts are orthogonal. Later Rao (1950) gives an elegant method of construction of fractional factorial experiments in blocks from orthogonal arrays when  $s$  is a prime or a prime power.

### 3. ASYMMETRICAL FACTORIAL EXPERIMENT

An experiment involving  $m$  factors  $F_1, F_2, \dots, F_m$  occurring at  $s_1, s_2, \dots, s_m$  levels respectively is called an asymmetrical factorial experiment, provided not all  $s_i$ 's are equal.

But the situation that occurs most often in practice is that there are  $g$  groups of factors—there being  $m_i$  factors in the  $i$ -th group, each occurring at  $s_i$  levels ( $i = 1, 2, \dots, g$ ;  $\sum_{i=1}^g m_i = m$ ). So the experiment  $s_1^{m_1} \times s_2^{m_2} \times \dots \times s_g^{m_g}$  may be regarded as a compound of  $g$  symmetrical factorial experiments. The problem of finding a suitable sub-set of the assemblies of the complete experiment, which preserves interactions upto desired order is important here for the same reasons as in a symmetrical factorial experiment and besides, it has an added interest because of its general nature.

A solution to the above problem is provided by the following.

**Theorem 1:** *If orthogonal arrays  $(N_i, m_i, s_i, d_i + k_i - 1, \lambda_i)$   $i = 1, 2, \dots, g$  exist, then there exists an array which yields a fractional replicate design in  $\prod_{i=1}^g N_i = \prod_{i=1}^g \lambda_i s_i^{d_i + k_i - 1}$  assemblies of the asymmetrical factorial experiment  $s_1^{m_1} \times \dots \times s_g^{m_g}$ . Further, if  $i$ -th orthogonal array ( $i = 1, 2, \dots, g$ ) preserves all main effects and interactions of order upto  $(k_i - 1)$  on the assumption that interactions of order  $(d_i - 1)$  ( $d_i > k_i$ ) and higher are absent, then in the derived array, all main effects and interactions involving  $r = \sum_{i=1}^g r_i$  factors ( $0 < r \leq gk, 0 \leq r_i \leq k$ ) become measurable, where  $r_1$  factors are chosen from the first group of  $m_1$  factors,  $r_2$  factors from the second group and so on.*

**Proof:** The theorem will be proved for the case  $g = 2$ . The extension for any integer  $g > 2$  is almost immediate. Consider the two orthogonal arrays  $(N_1, m_1, s_1, d_1 + k_1 - 1, \lambda_1)$ ,  $(N_2, m_2, s_2, d_2 + k_2 - 1, \lambda_2)$ . A column is taken from the first array and just below it is put a column from the second array. As there are  $N_1$  columns in the first array and  $N_2$  columns in the second, the above method of combining the columns of the two arrays will generate  $N_1 N_2$  columns with no combination repeated and each of these new columns will have  $(m_1 + m_2)$  rows. In this new array, any combination of levels of  $t_1$  factors of the first group and  $t_2$  factors from the second group will be repeated in  $\lambda_1 s_1^{d_1 + k_1 - 1 - t_1} \lambda_2 s_2^{d_2 + k_2 - 1 - t_2}$  columns  $\{t_1 \leq d_1 + k_1 - 1; t_2 \leq d_2 + k_2 - 1\}$ .



Let  $F_1, F_2, \dots, F_{m_1}$  represent the  $m_1$  factors of the first group and  $G_1, G_2, \dots, G_{m_2}$  represent the  $m_2$  factors of the second group. Defining symbolically

$$F_i^a = m_{0a} F_{i0} + \dots + m_{is_1-1a} F_{is_1-1} \quad (i = 1, 2, \dots, m_1)$$

$$G_j^a = l_{0a} G_{j0} + \dots + l_{js_2-1a} G_{js_2-1} \quad (j = 1, 2, \dots, m_2)$$

$$\sum m_{ra} = 0 \quad \text{when } a \neq 0$$

$$\sum m_{ra} m_{rb} = 0 \quad \text{when } a \neq b$$

$$m_{ra} = 1 \quad \text{for all } r \quad \text{when } a = 0$$

$$\sum l_{ua} = 0 \quad \text{when } a \neq 0$$

$$\sum l_{ua} l_{ub} = 0 \quad \text{when } a \neq b$$

$$l_{ua} = 1 \quad \text{for all } u \quad \text{when } a = 0.$$

$F_{i0}, F_{i1}, \dots, F_{is_1-1}$  and  $G_{j0}, G_{j1}, \dots, G_{js_2-1}$  representing the levels of  $F_i$  and  $G_j$  respectively. Then the symbolic product  $F_1^a F_2^b \dots F_{m_1}^k G_1^a G_2^b \dots G_{m_2}^p$  may be taken to represent the interaction  $[a, b, \dots; \alpha, \beta, \dots]$  of the factors for which the values are not zero. The expression obtained from  $[a, b, \dots; \alpha, \beta, \dots]$  by retaining only the assemblies occurring in the derived array may be denoted by  $\{a, b, \dots; \alpha, \beta, \dots\}$  and a set of necessary and sufficient conditions (Rao, 1947) that this will measure the corresponding interaction is that it is not orthogonal to  $[a, b, \dots; \alpha, \beta, \dots]$  but is orthogonal to every other function of this type including the interactions which are absent.

Consider an expression  $\{a_1, a_2, \dots, a_{r_1}, 0, \dots, 0; \alpha_1, \alpha_2, \dots, \alpha_{r_2}, 0, \dots, 0\}$  where  $a_i \neq 0$  ( $i = 1, 2, \dots, r_1$ ),  $\alpha_j \neq 0$  ( $j = 1, 2, \dots, r_2$ ) for  $0 \leq r_1 \leq k_1, 0 \leq r_2 \leq k_2$  and  $0 < r_1 + r_2 \leq 2k$ . This is evidently a contrast and this is not orthogonal to  $[a_1, a_2, \dots, a_{r_1}, 0, \dots, 0; \alpha_1, \alpha_2, \dots, \alpha_{r_2}, 0, \dots, 0]$ .

Let in the expression  $\{a, b, \dots; \alpha, \beta, \dots\}$  there be  $t_1$  non-zero coordinates among the first  $m_1$  and  $t_2$  non zero coordinates among the second  $m_2$  and consider the case when this has got no non-zero coordinate in common with those of  $\{a_1, a_2, \dots, a_{r_1}, 0, \dots, 0; \alpha_1, \alpha_2, \dots, \alpha_{r_2}, 0, \dots, 0\}$ . Then, since any assembly of a given set of  $(r_1 + t_1) \leq d_1 + k_1 - 1$  factors from the first group will be repeated the same number of times with all the assemblies of another given set of  $(r_2 + t_2) \leq d_2 + k_2 - 1$  factors from the second group, it follows that  $\{a_1, a_2, \dots, a_{r_1}, 0, \dots, 0; \alpha_1, \alpha_2, \dots, \alpha_{r_2}, 0, \dots, 0\}$  is orthogonal to  $\{a, b, \dots; \alpha, \beta, \dots\}$  and hence to  $[a, b, \dots; \alpha, \beta, \dots]$ . When  $\{a, b, \dots; \alpha, \beta, \dots\}$  has some non-zero coordinates in common with those of  $\{a_1, a_2, \dots, a_{r_1}, 0, \dots, 0; \alpha_1, \alpha_2, \dots, \alpha_{r_2}, 0, \dots, 0\}$  it follows from similar considerations as above, that  $\{a_1, a_2, \dots, a_{r_1}, 0, \dots, 0; \alpha_1, \alpha_2, \dots, \alpha_{r_2}, 0, \dots, 0\}$  is orthogonal to  $[a, b, \dots; \alpha, \beta, \dots]$ . So  $\{a_1, a_2, \dots, a_{r_1}, 0, \dots, 0; \alpha_1, \alpha_2, \dots, \alpha_{r_2}, 0, \dots, 0\}$  by Rao's theorem, defines the corresponding interaction involving the factors  $F_1, F_2, \dots, F_{r_1}$  of first group and factors  $G_1, G_2, \dots, G_{r_2}$  from the second group. This establishes the theorem for the case  $g = 2$ .

# FRACTIONAL REPLICATES AND PARTIALLY BALANCED ARRAYS

If now, there exists an orthogonal array  $(N_3, m_3, s_3, d_3 + k_3 - 1, \lambda_3)$  then by taking a column from the derived matrix and putting below it a column from the third orthogonal array and repeating this operation so that no combination of any two columns is repeated we get an array in  $N_1 N_2 N_3$  columns and  $(m_1 + m_2 + m_3)$  rows. And thus taking a new orthogonal array at each stage and combining the columns of these with those of the derived matrix obtained at the preceding stage, we will finally get an array in  $N_1 N_2 \dots N_g$  columns and  $m_1 + m_2 + \dots + m_g$  rows with the stated properties.

To illustrate the general theorem proved above, let us consider an

*Example :* Fractional Replication in  $2^2.3^2 = 36$  assemblies of a  $2^3.3^4$  experiment.

It is known that the hypercubes  $(2^2, 3, 2, 2)$  and  $(3^2, 4, 3, 2)$  exist and each one of them accommodates maximum number of factors and preserves main effects on the assumption that higher order interactions are absent. The arrays are

TABLE 1. ARRAY:  $(2^2, 3, 2, 2)$

factors	assemblies			
	1	2	3	4
$F_1$	0	1	1	0
$F_2$	0	1	0	1
$F_3$	0	0	1	1

ARRAY:  $(3^2, 4, 2, 2)$

factors	assemblies								
	1	2	3	4	5	6	7	8	9
$G_1$	0	0	0	1	1	1	2	2	2
$G_2$	0	1	2	1	2	0	2	0	1
$G_3$	0	2	1	1	0	2	2	1	0
$G_4$	0	1	2	0	1	2	0	1	2

TABLE 2. DERIVED ARRAY:  $(2^2, 3, 2, 2) \times (3^2, 4, 2, 2)$

factors	assemblies											
	1	2	3	4	5	6	7	8	9	10	11	12
$F_1$	0	1	1	0	0	1	1	0	0	1	1	0
$F_2$	0	1	0	1	0	1	0	1	0	1	0	1
$F_3$	0	0	1	1	0	0	1	1	0	0	1	1
$G_1$	0	0	0	0	0	0	0	0	0	0	0	0
$G_2$	0	0	0	0	1	1	1	1	2	2	2	2
$G_3$	0	0	0	0	2	2	2	2	1	1	1	1
$G_4$	0	0	0	0	1	1	1	1	2	2	2	2

assemblies (continued)												
factors	13	14	15	16	17	18	19	20	21	22	23	24
$F_1$	0	1	1	0	0	1	1	0	0	1	1	0
$F_2$	0	1	0	1	0	1	0	1	0	1	0	1
$F_3$	0	0	1	1	0	0	1	1	0	0	1	1
$G_1$	1	1	1	1	1	1	1	1	1	1	1	1
$G_2$	1	1	1	1	2	2	2	2	0	0	0	0
$G_3$	1	1	1	1	0	0	0	0	2	2	2	2
$G_4$	0	0	0	0	1	1	1	1	2	2	2	2

assemblies (continued)												
factors	25	26	27	28	29	30	31	32	33	34	35	36
$F_1$	0	1	1	0	0	1	1	0	0	1	1	0
$F_2$	0	1	0	1	0	1	0	1	0	1	0	1
$F_3$	0	0	1	1	0	0	1	1	0	0	1	1
$G_1$	2	2	2	2	2	2	2	2	2	2	2	2
$G_2$	2	2	2	2	0	0	0	0	1	1	1	1
$G_3$	2	2	2	2	1	1	1	1	0	0	0	0
$G_4$	0	0	0	0	1	1	1	1	2	2	2	2

This design in 36 experimental units will allow estimation of all the main effects on the assumption that interactions involving any two members or more of the same group are absent and besides, the first order interactions involving any member of the first group and any member of the second will also become measurable.

The analysis of variance table may be set up as follows:

TABLE 3. ANALYSIS OF VARIANCE

degrees of freedom			
main effects	$F$	$3(2-1) = 3$	
	$G$	$4(3-1) = 8$	
first order interaction	$FG$	$3 \times 4(2-1)(3-1) = 24$	
total	...	...	35

## FRACTIONAL REPLICATES AND PARTIALLY BALANCED ARRAYS

Then, this design has no degrees of freedom left for estimation of error. If the first order interactions  $FG$  are absent then these will provide an estimate of error. Sometimes estimates of error variance based on previous experience may be available and these may be used in such situations. When, however, such estimates are not available and there are no *a priori* reasons to assume that interactions  $FG$  will not produce any effect we may repeat the fractional replicate designs to get an estimate of error variance.

Corollary to Theorem 1: If  $s_1, s_2, \dots, s_g$  are primes or powers of primes, then the hypercubes  $\left( s_i^{t_i}, n_i = \frac{s_i^{t_i} - 1}{s_i - 1}, s_i, 2 \right) i = 1, 2, \dots, g$  can be combined in the manner of Theorem 1 to get a fractional replicate design for the asymmetrical factorial experiment  $s_1^{n_1} \times s_2^{n_2} \times \dots \times s_g^{n_g}$  and this design will preserve (i) all main effects on the assumption that interactions involving two factors or more from the same group are absent, (ii) all interactions involving upto  $g$  factors but no two factors present in the interaction should come from the same group.

*Proof:* If  $s_1, s_2, \dots, s_g$  are primes or powers of primes, then according to Rao's theorem (1946) the hypercubes  $\left( s_i^{t_i}, n_i = \frac{s_i^{t_i} - 1}{s_i - 1}, s_i, 2 \right)$  exist and the  $i$ -th hypercube in  $s_i^{t_i}$  assemblies accommodate the maximum number of factors

$$n_i = \frac{s_i^{t_i} - 1}{s_i - 1} \quad (i = 1, 2, \dots, g).$$

Further, all the  $n_i$  main effects from the  $i$ -th hypercube are estimable on the assumption that interaction of order  $d \geq 1$  are absent and since each one of them carries  $(s_i - 1)$  degrees of freedom, these exhaust the  $(s_i^{t_i} - 1)$  degrees of freedom associated with the  $i$ -th hypercube in  $s_i^{t_i}$  assemblies. When these hypercubes are combined in the manner of Theorem 1, an array in  $s_1^{t_1} \times s_2^{t_2} \times \dots \times s_g^{t_g}$  assemblies is obtained and it follows from Theorem 1 that this will have the properties as stated in the enunciation.

Example 1 illustrates the corollary just proved, for the case  $g = 2$ .  $s_1 = 2, t_1 = 2, s_2 = 3$  and  $t_2 = 2$ .

### 4. CONSTRUCTION OF SOME IMPORTANT DESIGNS

Using the theorem proved above, it is now possible to construct fractional replicate designs for asymmetrical factorial experiments from the orthogonal arrays already constructed and listed in Rao (1946, 1947). Plackett and Burman (1946). Bose and Bush (1952), Bush (1952). A list of designs for some of the important experiments likely to occur in practice is given here. In a later section, two other methods of construction of fractional replicate designs for asymmetrical factorial



experiments will be described. These methods are useful only for limited values of  $g$ 's and  $s_i$ 's. But such designs effect a saving in the number of experimental units, otherwise required by the general method described above.

The class of experiments  $2^{\lambda-1} . s$ , where  $\lambda$  a positive integer, requires  $4\lambda s$  assemblies for a fractional replicate design which preserves all the  $4\lambda$  main effects and all  $(4\lambda-1)$  first order interactions between any member of the first group and the factor with  $s$  levels. Since orthogonal arrays of strength 2 for  $s=2$  and  $\lambda=25$  are available in Plackett and Burman (1946), such designs can be easily constructed. Experiments  $2^k . s$ ,  $4\lambda-4 \leq k < 4\lambda-1$  also require  $4\lambda s$  assemblies for a fractional replicate design.

TABLE 4. LIST OF SOME IMPORTANT FRACTIONAL REPLICATE DESIGNS FOR ASYMMETRICAL FACTORIAL EXPERIMENTS

(designs based on orthogonal arrays of strength 2)

description of the complete factorial experiment	no. of assemblies required	orthogonal arrays which combine to give the design	nature of effects measurable
(1)	(2)	(3)	(4)
1. $2^k . s$ $(4\lambda-5 < k \leq 4\lambda-1)$	$4\lambda s$	$(4\lambda k, 2, 2, \lambda) \times (s)$	all main effects and interactions involving two factors one from each group. (On the assumption interaction of order $d \geq 1$ within each group absent.)
2. $2^k . 3^l$ $\left[ \begin{array}{l} 4\lambda-5 < k \leq 4\lambda-1, \\ 3^{l-1} < 2l+1 \leq 3^l \end{array} \right]$	$4\lambda . 3^l$	$(4\lambda, k, 2, 2, \lambda)$ $\times$ $(3^l, l, 3, 2, 3^{l-2})$	all main effects and interactions involving two factors one from each group. (On the assumption interactions of order $d \geq 1$ within each group absent.)
3. $2^k . 4^m$ $\left[ \begin{array}{l} 4\lambda-5 < k \leq 4\lambda-1 \\ 4^{l-1} < 3m+1 \leq 4^l \end{array} \right]$	$4\lambda . 4^l$	$(4\lambda, k, 2, 2, \lambda)$ $\times$ $(4^l, m, 4, 2, 4^{l-2})$	-do-
4. $2^k . 3^{l_1} . 4^{l_2}$ $\left[ \begin{array}{l} 4\lambda-5 < k \leq 4\lambda-1 \\ 3^{l_1-1} < 2l_1+1 \leq 3^{l_1} \\ 4^{l_2-1} < 3m+1 \leq 4^{l_2} \end{array} \right]$	$4\lambda . 3^{l_1} . 4^{l_2}$	$(4\lambda, k, 2, 2, \lambda)$ $\times$ $(3^{l_1}, l_1, 3, 2, 3^{l_1-2})$ $\times$ $(4^{l_2}, m, 4, 2, 4^{l_2-2})$	all main effects and all first order interactions and second order interactions involving not more than one from each group are estimable. (on the assumption interactions of order $d \geq 1$ within each group absent.)

In the above list only orthogonal arrays of strength 2, have been considered. Orthogonal arrays of strength higher than 2 may also be combined in a similar manner but the designs obtained from them will require comparatively large number of assemblies.

# FRACTIONAL REPLICATES AND PARTIALLY BALANCED ARRAYS

## 5. FRACTIONAL REPLICATE DESIGNS FOR $s_1'' . s_2$ EXPERIMENTS WHERE $s_2 = s_1'$ AND $s_1$ IS A PRIME OR POWER OF A PRIME

Fractional replicate designs for asymmetrical factorial experiments  $s_1'' . s_2$  where  $s_1$  is a prime or power of a prime and  $s_2 = s_1'$  can be constructed in considerably reduced number of assemblies. These designs are made available by a theorem (Rao, 1950): The necessary and sufficient conditions that the  $s'$  combinations obtained as a solution of the set of  $(n-r)$  independent homogeneous equations

$$a_{i1} x_1 + \dots + a_{in} x_n = 0 \quad i = 1, 2, \dots, n-r \quad \dots \quad (5.1)$$

define an array of strength  $(d+1)$  is that no equation derivable as a linear combination from the  $(n-r)$  equations determining the subset contains less than  $(d+2)$  non-null coefficients.

He, then, uses this theorem to construct block designs for fractional factorial experiments which preserve main effects and first order interactions on the assumption that interactions of order  $d-1$  or higher are absent.

For that, he proceeds as follows: Another set of  $t$  homogeneous equations (properly chosen)

$$a_{j1} x_1 + \dots + a_{jn} x_n = 0 \quad j = n-r+1, \dots, n-r+t \quad \dots \quad (5.2)$$

are considered together with (5.1), so that the  $s^{r-t}$  assemblies obtained as solutions of this set of  $(n-r+t)$  equations will define an array of strength 2. This set of  $s^{r-t}$  assemblies may be called the key array. The  $s'$  assemblies obtained as a solution of (5.1) are divisible into  $s^t$  such arrays each of strength 2. These are obtained by adding to each member of the key array, a solution of (5.1) which does not already occur in the key array. Now if there are  $n$  factors each at  $s_1$  levels and another factor at  $s_1'$  levels, then to each one of the  $s_1^{r-t}$  assemblies of a group is added one level of the additional factor. The  $s_1^t$  groups will each have a different level of the additional factor. Thus we have in  $s_1^r$  assemblies, a fractional replicate design of an asymmetrical factorial experiment  $s_1'' . s_2$  where  $s_2 = s_1'$  and  $s_1$  is a prime or power of a prime.

This design will allow estimation of all the  $(n+1)$  main effects, and  $\binom{n}{2}$  first order interactions between any two factors of the first group on the assumption that interactions involving  $d$  factors or more of the first group are absent. The analysis

of variance table may be set up as follows:

TABLE 5. ANALYSIS OF VARIANCE FOR FRACTIONAL REPLICATE DESIGN

factors		degrees of freedom
main effects	first group of factors	$n(s_1 - 1)$
	extra factor	$(s_2 - 1)$
first order interaction	between factors of first group	$\binom{n}{2} (s_1 - 1)^2$
error		(obtained by subtraction)
total		$s^r - 1$

A further reduction in the number of assemblies can be achieved if we are not interested in preserving all the  $\binom{n}{2}$  first order interactions between factors of the first group but only in a selected sub-set of first order interactions. Then each group of  $s_1^{r-t}$  assemblies need not be an orthogonal array of strength 2 but it will do simply if each group of  $s_1^{r-t}$  assemblies satisfies the following less restrictive properties

- (i) all the levels of any factor of first set occur equal number of times, this number being same for all factors and groups,
- (ii) all combinations of levels of any two factors whose interaction we want to preserve, should occur equal number of times, this number being same for all pairs and groups and all the  $s^r$  assemblies together should form an orthogonal array of strength  $d \geq 4$ .

Consider the factorial experiment  $2^5 \times 4$ . For five factors each at 2 levels, a hypercube of strength 4 in 16 assemblies exists and this allow estimation of main effects and first order interactions on the assumption that interactions of order  $d \geq 2$  are absent. Let us denote the first five factors by  $F_1, F_2, \dots, F_5$  and the one at 4 levels by  $G$ . Now, since each level of  $G$  is to occur equal number of times, in order to preserve all first order interactions of any two  $F$ 's it is necessary to construct an orthogonal array of strength 2 in 4 assemblies. But an orthogonal array of strength 2 in 4 assemblies can accommodate only upto three factors. And since 16 assemblies provide us with only 15 degrees of freedom and 8 degrees of freedom are taken up by the 6 main effects, only 7 degrees of freedom are left and we can at best estimate 7 first order interactions involving  $F$ 's alone. Then no degrees of freedom will be left for estimation of error variance. If a previous estimate of error variance is available then this need not worry us. The interactions to be preserved will be decided by the nature of the experiment. Having decided on the interactions to be preserved the problem is to divide the 16 assemblies into four groups so that conditions (i) and (ii) are satisfied.

*Example 2:* An example of this experiment taken from Davies and Hay (1950) is given below. The experiment concerned investigation of effects of several factors on yield of penicillin. Five factors  $A, B, C, D, E$  each at two levels and a factor



# FRACTIONAL REPLICATES AND PARTIALLY BALANCED ARRAYS

$F$  at four levels were chosen. The factors were

TABLE 6. DESCRIPTION OF EXPERIMENTAL STAGES

stage 1	preparation of inoculum
$A$	concentration of corn steep liquor
$B$	amount of sugars
$C$	quality of sugars
stage 2	fermentation
$D$	concentration of corn steep liquor obtained from first stage.
$E$	quality of corn steep liquor
$F$	4 fermenters

A design in 16 assemblies preserving the main effects and the 7 first order interactions  $AB, AC, AD, AE, BD, CD$  and  $DE$  was adopted and the detailed analysis is also given in Davies and Hay (1950), Davies (1954). They have constructed the design from different considerations. We have constructed the design with the help of orthogonal arrays and in the manner described in an earlier paragraph. The design happens to be the same as that obtained by Davies and Hay. The design is given below:

TABLE 7. DESIGN OF THE EXPERIMENT

factors	assemblies															
	first group				second group				third group				fourth group			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$A$	1	0	0	1	0	1	1	0	1	0	0	1	0	1	1	0
$B$	1	0	1	0	0	1	0	1	0	1	0	1	1	0	1	0
$C$	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0
$D$	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0
$E$	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1
$G$	0	0	0	0	1	1	1	1	2	2	2	2	3	3	3	3

The splitting up of degrees of freedom will be as follows:

TABLE 8. ANALYSIS OF VARIANCE

factors		degrees of freedom
main effects	$A$	1
	$B$	1
	$C$	1
	$D$	1
	$E$	1
	$G$	3
first order interactions	$AB$	1
	$AC$	1
	$AD$	1
	$AE$	1
	$BD$	1
	$CD$	1
	$DE$	1
total		15



6. DESIGNS FOR  $2 \times 3^k$  AND EXPERIMENTS IN STAGES

If in experiments  $2 \times 3^k$ , one is allowed to relax the condition that all the levels of a factor should occur equal number of times, fractional replicate designs for such experiments can be constructed by taking an orthogonal array with  $(k+1)$  factors each at three levels and then replacing the highest level of one factor by one of the two lower levels, all through the array. So that for this particular factor, one level will occur twice as many times as the other. This method of construction sometimes effects a saving in the number of experimental units to be used e.g. in the case of orthogonal arrays of strength 2, if  $2k+1 > 3^{t-1}$  and  $2k-3 \leq 3^t$  where  $t$  is some positive integer.

Experiments in industries are often conducted in several stages and different number of factors have to be introduced at different stages. The total number of units to be used in the experiment is sometimes decided by the varying costs of setting up different stages of the experiment. If products of an earlier stage, which have to be used in a later stage, are costly, we may have to be content only with the estimates of main effects of the factors introduced at the earlier stage, while possibly for the factors introduced at a later stage, it may be possible to estimate interactions upto a certain order. This restriction on the experiment may be taken care of, in the design in the case of a symmetrical factorial experiment by taking an orthogonal array of lower strength for the factors of the earlier stage and an orthogonal array of higher strength for the factors of the later stage and combining the two arrays in a manner so that the derived array remains an orthogonal array. Or alternatively, starting with an orthogonal array of suitable strength for all the factors of an experiment, it is sometimes possible, to find out within the array a sub-group of factors which amongst themselves form an orthogonal array of lower strength say 2, while the remaining factors form one of a higher strength. Now, if each of the distinct sub-assemblies of the factors of the sub-group forming an array of strength 2 occurs more than once, then these may serve as factors of the earlier stage of the experiment while the rest of the factors may be introduced in the later stage of the experiment. As an example, we may consider the following array.

TABLE 9. ARRAY: (23, 7, 2, 2,)

factors	assemblies							
	1	2	3	4	5	6	7	8
A	1	1	1	1	0	0	0	0
B	1	1	0	0	1	1	0	0
C	1	0	1	0	1	0	1	0
D	1	0	0	1	0	1	1	0
E	1	1	0	0	0	0	1	1
F	1	0	1	0	0	1	0	1
G	1	0	0	1	1	0	0	1

## FRACTIONAL REPLICATES AND PARTIALLY BALANCED ARRAYS

Here the factors  $E, F, G$  form an array of strength 2 and each of 4 distinct sub-assemblies occur twice while the four factors  $A, B, C, D$  form an array of strength 3. So in an experiment which has to be conducted in two stages and where it is costly to have many experimental units in the first stage, the factors  $E, F, G$  may be used at the first stage which will require only 4 units and the factors  $A, B, C, D$  may be introduced in the following stage. Examples of experimental situations where the problems discussed above, have occurred are provided by Taguchi (1955). On behalf of the Quality Control Unit of Indian Statistical Institute, he had conducted several experiments in different industries in South India. His reports submitted to the Indian Statistical Institute contain description of designs, analysis and inferences drawn. But he does not discuss how he constructed these designs. A brief description of one of the experiments together with the design adopted by him is given here to illustrate the applicability of the designs constructed in this paper.

*Experiment:* A  $2 \times 3^7$  experiment to find an optimum operational standard for anodising aluminium alloy parts: All aluminium base alloy parts have to be given an anodic treatment for forming a thin film of oxide coating on the metal by an electrolyte oxidation process. The problem was to design a suitable experiment in two stages, four factors to be introduced in one stage and four factors in the other stage; all the factors chosen excepting one, were at 3 levels. The first stage consists in preparation of an anodising bath which is controlled by the four factors.

TABLE 10. FACTORS OF THE FIRST STAGE

		levels		
C	concentration of bath (acid content)	50 gms/litre	(1)	
		40 gms/litre	(2)	
		30 gms/litre	(3)	
D	voltage cycle	38V-50V	(1)	
		40V-50V	(2)	
		45V-52V	(3)	
E	time cycle	D(1)	D(2)	D(3)
		(1) 10-30-5-5 (50m)	10-25-5-5 (45m)	10-20-4-4 (38m)
		(2) 10-35-5-5 (55m)	10-30-5-5 (50m)	10-25-4-4 (43m)
		(3) 10-40-5-5 (60m)	10-35-5-5 (55m)	10-30-4-4 (48m)
F	temperature of bath	(1) 100°F		
		(2) 106°F		
		(3) 103°F		

The symbol 10-30-5-5 (50m) under  $D(1)$  implies that the voltage has to be raised from 0 to 38 volts in 10 minutes, maintained at 38 volts for 30 minutes and again raised to 50 volts in 5 minutes and kept at 50 volts for 5 minutes—total time taken for the entire operation being 50 minutes. The second stage of the experiment concerns

the alloy parts that have to be suspended from the anode into the bath by pure aluminium. Factors which control this stage of the experiment are

TABLE 11. FACTORS OF THE SECOND STAGE

		levels	
<i>A</i>	degreasing operation time	8 minutes	(1)
		10 minutes	(2)
		12 minutes	(3)
<i>B</i>	type of parts	big	(1)
		medium	(2)
		small	(3)
<i>G</i>	rinsing time in cold water swill	2 minutes	(1)
		3 minutes	(2)
<i>H</i>	rinsing time in hot water bath (150°F)	3 minutes	(1)
		4 minutes	(2)
		5 minutes	(3)

Cost and difficulty of operation limit the number of baths that can be set up. The design has to take care of this feature of the experiment. So the different sub-assemblies of the factors *C, D, E, F* which we can introduce in the experiment are limited while we may have sufficiently large number of sub-assemblies for *A, B, G, H*. Taguchi's design is given below:

TABLE 12. DESIGN FOR  $2 \times 3^7$  EXPERIMENT

TABLE 12. DESIGN FOR 2x3 <sup>7</sup> EXPERIMENT									
assemblies	factors								
	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>		<i>B</i>	<i>A</i>	<i>G</i>	<i>H</i>
1-3	3	2	1	1	—	2 1 3	3 1 2	1 1 2	3 2 1
4-6	3	1	2	2	—	2 1 3	1 2 3	1 2 1	1 3 2
7-9	3	3	3	3	—	2 1 3	2 3 1	2 1 1	2 1 3
10-12	2	3	2	1		2 1 3	3 1 2	1 2 1	1 3 2
13-15	2	2	3	2	—	2 1 3	1 2 3	2 1 1	2 1 3
16-18	2	1	1	3	—	2 1 3	2 3 1	1 1 2	3 2 1
19-21	1	1	3	1	—	2 1 3	3 1 2	2 1 1	2 1 3
22-24	1	3	1	2	—	2 1 3	1 2 3	1 1 2	3 2 1
25-27	1	2	2	3		2 1 3	2 3 1	1 2 1	1 3 2



## FRACTIONAL REPLICATES AND PARTIALLY BALANCED ARRAYS

If we suppose that  $G$  had also three levels to start with (and later 3 has been converted into 1), then it is easily seen that the 27 sub-assemblies of  $A, B, G, H$  form an orthogonal array of strength 3. Further, each of the groups of the sub-assemblies *viz.*, 1-9, 10-18, 19-27 form an orthogonal array of strength 2. The 9 distinct sub-assemblies involving  $C, D, E, F$  alone, define an orthogonal array of strength 2 and each of these sub-assemblies was tagged on to three sub-assemblies involving  $A, B, G, H$  so that this becomes an orthogonal array of strength 2 in 27 assemblies involving 8 factors. It is further seen that the interaction  $BC$  is preserved, the condition for which given by Rao. (1947), states that all combinations of 3 factors involving both  $B$  and  $C$  should occur an equal number of times in an orthogonal array ( $N, n, s, 2$ ). Taguchi (1955) has not discussed the features of the design, that are given above. The splitting up of degrees of freedom was done as follows:

TABLE 13. ANALYSIS OF VARIANCE

	factors	degrees of freedom
main effects	$C$	2
	$D$	2
	$E$	2
	$F$	2
	$B$	2
	$A$	2
	$G$	1
	$H$	2
interaction	$BC$	4
	error	7
	total	26

### 7. ANALYSIS OF FRACTIONAL REPLICATE DESIGNS

Systematic methods (Yates' technique) of analysis of fractional replicate designs are available in the case of symmetrical factorial experiments (Davies 1954). If, however, the interest is mainly on main effects and first order interactions, systematic methods may prove to be laborious and it may be advantageous to obtain estimates of main effects and lower order interactions by writing down the contrasts omitting those assemblies which do not occur in the array. A great simplicity results in the analysis of designs if for any two mutually orthogonal estimable linear functions of parameters, the corresponding best estimates are uncorrelated.

If we consider the set up

$$E(y) = \tau A'$$

where  $y$  is the vector of  $n$  observations,  $\tau$  is a vector of  $k$  parameters and  $A'$  is the design matrix of the form  $k \times n$  and of rank  $r$ , then a linear function  $\tau p'$  is estimable if there exists an  $l$  such that

$$l A' A = p \quad (\text{Rao, 1952})$$



**Theorem:** *The necessary and sufficient condition that the best linear estimates of two estimable linear functions  $\tau p'_1$  and  $\tau p'_2$  ( $p_1 p'_2 = 0$ ), are uncorrelated is that*

$$(A' A) \cdot (A' A) = \mu \cdot A' A$$

*i.e. that non-zero roots of  $A' A$  are all equal.*

*Proof:* Sufficiency: Since  $\tau p'_1$  and  $\tau p'_2$  are estimable, there exist  $l_1$  and  $l_2$  so that

$$l_1 A' A = p_1$$

and

$$l_2 A' A = p_2$$

and covariance of the best estimates  $l_1 Q'$  and  $l_2 Q'$  is given by

$$l_1 A' A l'_2 \sigma^2$$

where  $Q = yA$ .

Now

$$p_1 p'_2 = l_1 A' A \cdot A' A l'_2 = 0. \quad \dots (7.1)$$

So if

$$A' A \cdot A' A = \mu A' A \quad \dots (7.2)$$

then (7.1) would imply that

$$l_1 A' A l'_2 = 0 \quad \dots (7.3)$$

*Necessity:* To prove that the condition is necessary, we will have to show that

$$\text{if } p_1 p'_2 = 0 \quad \dots (7.4)$$

$$\text{implies } l_1 A' A l'_2 = 0 \quad \dots (7.5)$$

then

$$(A' A)^2 = \mu A' A$$

or the non-zero roots of  $A' A$  are all equal.

In order that  $p_1 \tau'$  and  $p_2 \tau'$  are estimable, it is known that

$$l_1 A' A = p_1 = \alpha_1 C_1 + \alpha_2 C_2 + \dots + \alpha_r C_r \quad \dots (7.6)$$

$$l_2 A' A = p_2 = \alpha'_1 C_1 + \alpha'_2 C_2 + \dots + \alpha'_r C_r \quad \dots (7.7)$$

where  $C_1, C_2, \dots, C_r$  are the normalized latent vectors of  $A' A$  corresponding to its  $r$  positive latent-roots  $\lambda_1, \lambda_2, \dots, \lambda_r$  and  $(\alpha_1, \alpha_2, \dots, \alpha_r)$  and  $(\alpha'_1, \alpha'_2, \dots, \alpha'_r)$  are arbitrary constants.

We get from (7.4).

$$\alpha_1 \alpha'_1 + \alpha_2 \alpha'_2 + \dots + \alpha_r \alpha'_r = 0. \quad \dots (7.8)$$

# FRACTIONAL REPLICATES AND PARTIALLY BALANCED ARRAYS

Now, since  $C_i A' A = \lambda_i C_i$  ( $i = 1, 2, \dots, r$ ) from (7.5), (7.6) and (7.7) we get

$$\begin{aligned} l_1 A' A l'_2 &= (\alpha_1 C_1 + \dots + \alpha_r C_r) l'_2 \\ &= \left( \frac{\alpha_1}{\lambda_1} C_1 A' A + \dots + \frac{\alpha_r}{\lambda_r} C_r A' A \right) l'_2 \\ &= \left( \frac{\alpha_1}{\lambda_1} C_1 + \dots + \frac{\alpha_r}{\lambda_r} C_r \right) \cdot A' A l'_2 \\ &= \left( \frac{\alpha_1}{\lambda_1} C_1 + \dots + \frac{\alpha_r}{\lambda_r} C_r \right) (\alpha'_1 C_1 + \alpha'_2 C_2 + \dots + \alpha'_r C_r)' \\ &= \frac{\alpha_1 \alpha'_1}{\lambda_1} + \dots + \frac{\alpha_r \alpha'_r}{\lambda_r} = 0 \quad \dots (7.9) \end{aligned}$$

Since  $\alpha_i$ 's and  $\alpha'_i$ 's can be arbitrarily fixed subject to (7.8), it follows that  $\lambda_1 = \lambda_2 = \dots = \lambda_r = \mu$

Hence the theorem.

Analysis of a fractional replicate design of an asymmetrical factorial experiment is given here.

Analysis of the derived array  $(2^2, 2, 2, 2) \times (3^2, 4, 3, 2)$ : An artificial example constructed using the above design for an asymmetrical factorial experiment  $2^3 \times 3^4$  is analysed here. Since this design preserves the main effects and interactions involving two factors—only one factor being taken from one group and as their estimates are uncorrelated, the analysis is easily done by forming  $2 \times 3$  tables like,

TABLE 14. $F_1 \times G_1$				
	levels of $G_1$			total
	0	1	2	
levels of $F_1$	0			$F_{01}$
	1			$F_{11}$
	total	$G_{01}$	$G_{11}$	$G_{12}$
				$T$

Then, sum of squares due to  $F_1 = \frac{F_{01}^2 + F_{11}^2}{18} - \frac{T^2}{36}$

Sum of squares due to  $G_1 = \frac{G_{01}^2 + G_{11}^2 + G_{12}^2}{12} - \frac{T^2}{36}$

Sum of squares due to  
interaction  $F_1 \cdot G_1 = \text{Total corrected s.s. due to the table } F_1 \times G_1$   
—s.s. due to  $F_1$ —s.s. due to  $G_1$

and similar computations for 11 other tables will complete the analysis which is given below. Since we do not have any degrees of freedom left for estimation of error, tests of significance for the main effects may be performed on the assumption that  $FG$  interactions are absent.

TABLE 15. DESIGN AND YIELD OF THE EXPERIMENT

sl. no.	design							yield <i>y</i>
	$G_4$	$G_3$	$G_2$	$G_1$	$F_3$	$F_2$	$F_1$	
1	0	0	0	0	0	0	0	48
2	0	0	0	0	0	1	1	53
3	0	0	0	0	1	1	0	65
4	0	0	0	0	1	0	1	50
5	0	1	1	1	0	0	0	67
6	0	1	1	1	0	1	1	74
7	0	1	1	1	1	1	0	75
8	0	1	1	1	1	0	1	74
9	1	2	1	0	0	0	0	72
10	1	2	1	0	0	1	1	76
11	1	2	1	0	1	1	0	78
12	1	2	1	0	1	0	1	84
13	1	0	2	1	0	0	0	72
14	1	0	2	1	0	1	1	73
15	1	0	2	1	1	1	0	82
16	1	0	2	1	1	0	1	76
17	1	1	0	2	0	0	0	62
18	1	1	0	2	0	1	1	79
19	1	1	0	2	1	1	0	78
20	1	1	0	2	1	0	1	60
21	2	1	2	0	0	0	0	95
22	2	1	2	0	0	1	1	87
23	2	1	2	0	1	1	0	88
24	2	1	2	0	1	0	1	81
25	2	2	0	1	0	0	0	85
26	2	2	0	1	0	1	1	85
27	2	2	0	1	1	1	0	82
28	2	2	0	1	1	0	1	77
29	2	0	1	2	0	0	0	72
30	2	0	1	2	0	1	1	80
31	2	0	1	2	1	1	0	81
32	2	0	1	2	1	0	1	80
33	0	2	2	2	0	0	0	84
34	0	2	2	2	0	1	1	79
35	0	2	2	2	1	1	0	88
36	0	2	2	2	1	0	1	88

# FRACTIONAL REPLICATES AND PARTIALLY BALANCED ARRAYS

TABLE 16. ANALYSIS OF VARIANCE

factors		d.f.	s.s.	mean square	F
main effects	$F_1$	1	2.25	2.25	
	$F_2$	1	124.69	124.69	5.08*
	$F_3$	1	78.03	78.03	3.18
	$G_1$	2	175.50	87.75	3.58*
	$G_2$	2	1066.67	533.33	21.74**
	$G_3$	2	920.16	460.08	18.76**
	$G_4$	2	930.66	465.33	18.97**
interaction	$F_1G_1$	2	26.16		
	$F_1G_2$	2	97.99		
	$F_1G_3$	2	3.17		
	$F_1G_4$	2	32.67		
	$F_2G_1$	2	7.72		
	$F_2G_2$	2	110.89		
	$F_2G_3$	2	74.40		
	$F_2G_4$	2	13.56		
	$F_3G_1$	2	14.38		
	$F_3G_2$	2	22.88		
	$F_3G_3$	2	51.39		
	$F_3G_4$	2	133.56		
	sub-total for interactions	24	588.77		
	total	35	3886.73		

Main effects  $G_2$ ,  $G_3$  and  $G_4$  are significant at both the levels 5% and 1% while  $F_2$  and  $G_1$  are significant at 5% only.  $F_1$  and  $F_3$  are not significant at all.

## 8. PARTIALLY BALANCED ARRAYS

In an orthogonal array ( $s^t, n, s, 2$ ) where  $s$  is a prime or power of a prime, it is known (Rao 1946) that it can accommodate upto  $n = \frac{s^t-1}{s-1}$  factors. In this case, all the  $n$  main effects take up the  $(s^t-1)$  degrees of freedom and if an estimate of error variance based on previous experience is available this may be considered as the most economic design. But if  $\frac{s^{t-1}-1}{s-1} < n < \frac{s^t-1}{s-1}$  then also we have to use an orthogonal array in  $s^t$  assemblies. If our interest is in the estimation of main effects only, then the observations  $s^t - n(s-1) - 1$  may be regarded as unnecessary. It is possible to economise the number of observations in such cases if we relax the conditions of an orthogonal array, to certain extent. But this will introduce a little more complication in the analysis of the design.

An array involving  $n$  factors  $F_1, F_2, \dots, F_n$  each with  $s$  levels will be called a *partially balanced array* of strength  $d$  if for any group of  $d$  factors ( $d \leq n$ ) a combination of levels of  $d$  factors,  $F_{1i_1}, F_{2i_2}, \dots, F_{di_d}$  occurs  $\lambda_{i_1 i_2 \dots i_d}$  times, where  $\lambda_{i_1 i_2 \dots i_d}$  remains the same for all permutations of a given set  $(i_1, i_2, \dots, i_d)$  and for any group of  $d$  factors,  $i_j$  ranging from 0 to  $s-1$  for all  $j$ . Then, it is obvious that this property holds also for any  $k \leq d$ . Let amongst the  $d$  integers  $(i_1, i_2, \dots, i_d)$ , 0 occur  $r_0$  times,



1 occur  $r_1$  times and so on;—or, in other words, let in the treatment combination  $F_{1i_1} F_{2i_2} \dots F_{di_d}$  there be  $r_0$  factors which occur at 0-th level,  $r_1$  factors which occur at level 1 and so on.

Then,

$$r_0 + r_1 + \dots + r_{s-1} = d$$

and each of the  $\frac{d!}{r_0! r_1! \dots r_{s-1}!}$  treatment combinations obtained by permuting  $(i_1, i_2, \dots, i_d)$  will have the same  $\lambda_{i_1 i_2 \dots i_d}$  attached to it. Value of  $\lambda_{i_1 i_2 \dots i_r}$ , where  $r < d$ , is easily obtained by summing  $\lambda_{i_1 i_2 \dots i_d}$  over  $(i_{r+1}, \dots, i_d)$  where each  $i$  of  $(i_{r+1}, \dots, i_d)$  ranges from 0 to  $s-1$ .

### 9. EXAMPLES OF PARTIALLY BALANCED ARRAYS AND ANALYSIS

Consider the following orthogonal array

TABLE 17. ARRAY:  $(2^3, 4, 2, 2)$

	assemblies							
	1	2	3	4	5	6	7	8
A	1	1	1	0	0	0	1	0
B	1	0	0	0	1	1	1	0
C	0	1	0	1	1	0	1	0
D	0	0	1	1	0	1	1	0

It is known that 4 factors each with 2 levels require 8 assemblies for constructing an orthogonal array of strength 2. If from the above array we omit assemblies 7 and 8, we get an arrangement in 6 assemblies, which has the properties that for any two factors say A, B the combination of levels of the type (0,0) or (1,1) occurs once while (1,0) or (0,1) occurs twice. So this satisfies the properties of a partially balanced array.

This design can be analysed by the method of least squares as follows. Minimising the expression

$$L = (y_1 - a_1 - b_1 - c_0 - d_0)^2 + \dots + (y_6 - a_0 - b_1 - c_0 - d_1)^2$$

Where  $y_i$  denotes the observation corresponding to the  $i$ -th assembly and  $a_i$  denotes the effect of  $i$ -th level of  $a$  and similarly for other constants, we get

$$3a_0 + (b_0 + 2b_1) + (c_0 + 2c_1) + (d_0 + 2d_1) = y_4 + y_5 + y_6 \quad \dots (9.1)$$

$$3a_1 + (2b_0 + b_1) + (2c_0 + c_1) + (2d_0 + d_1) = y_1 + y_2 + y_3 \quad \dots (9.2)$$

$$(a_0 + 2a_1) + 3b_0 + (c_0 + 2c_1) + (d_0 + 2d_1) = y_2 + y_3 + y_4 \quad \dots (9.3)$$

$$(2a_0 + a_1) + 3b_1 + (2c_0 + c_1) + (2d_0 + d_1) = y_1 + y_5 + y_6 \quad \dots (9.4)$$

$$(a_0 + 2a_1) + (b_0 + 2b_1) + 3c_0 + (d_0 + 2d_1) = y_1 + y_3 + y_6 \quad \dots (9.5)$$

$$(2a_0 + a_1) + (2b_0 + b_1) + 3c_1 + (2d_0 + d_1) = y_2 + y_4 + y_5 \quad \dots (9.6)$$

$$(a_0 + 2a_1) + (b_0 + 2b_1) + (c_0 + 2c_1) + 3d_0 = y_1 + y_2 + y_5 \quad \dots (9.7)$$

$$(2a_0 + a_1) + (2b_0 + b_1) + (2c_0 + c_1) + 3d_1 = y_3 + y_4 + y_6 \quad \dots (9.8)$$

# FRACTIONAL REPLICATES AND PARTIALLY BALANCED ARRAYS

Now taking one equation from each pair, and putting  $a_0 = 0$ ,  $b_0 = 0$ ,  $c_0 = 0$ ,  $d_0 = 0$ , we get

$$3a_1 + b_1 + c_1 + d_1 = y_1 + y_2 + y_3 = Q_1$$

$$a_1 + 3b_1 + c_1 + d_1 = y_1 + y_5 + y_6 = Q_2$$

$$a_1 + b_1 + 3c_1 + d_1 = y_2 + y_4 + y_5 = Q_3$$

$$a_1 + b_1 + c_1 + 3d_1 = y_3 + y_4 + y_6 = Q_4$$

Solving we get,

$$a_1 + b_1 + c_1 + d_1 = \frac{1}{3} \sum_{i=1}^6 y_i$$

$$\hat{a}_1 = \frac{1}{2} \left[ (y_1 + y_2 + y_3) - \frac{1}{3} \sum_{i=1}^6 y_i \right]$$

$$\hat{b}_1 = \frac{1}{2} \left[ (y_1 + y_5 + y_6) - \frac{1}{3} \sum_{i=1}^6 y_i \right]$$

$$\hat{c}_1 = \frac{1}{2} \left[ (y_2 + y_4 + y_5) - \frac{1}{3} \sum_{i=1}^6 y_i \right]$$

$$\hat{d}_1 = \frac{1}{2} \left[ (y_3 + y_4 + y_6) - \frac{1}{3} \sum_{i=1}^6 y_i \right]$$

Sum of squares due to a factor  $A$  is then, equal to  $\frac{12}{5} \hat{a}_1^2$ . Sum of squares due to main effects is  $\hat{a}_1 Q_1 + \hat{b}_1 Q_2 + \hat{c}_1 Q_3 + \hat{d}_1 Q_4$  and in this case we are left with 1 degree of freedom which may be used to get an estimate of error but this, however, may not be, very reliable.

In the general case of  $k$  factors each with  $s$  levels if we were trying out a partially balanced array of strength 2 the normal equations (on the assumption that all interaction are absent) may be seen to be

$$yA = (\tau_1 : \tau_2 : \dots : \tau_k) \begin{bmatrix} D & \Lambda & \Lambda & \dots & \Lambda \\ \dots & D & \Lambda & \dots & \Lambda \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & D \end{bmatrix}$$

where  $y$  is the vector of observations,  $\tau_1, \tau_2, \dots, \tau_k$  are the  $k$  groups of parameters,  $A$  is the transpose of the design matrix and in the partitioned matrix which is  $A'A$ ,  $D$  is a diagonal matrix with  $\mu_i$ 's in the diagonal line,

where  $\mu_i$  represents the number of times the  $i$ -th level of a factor occurs in the array and

$$\Lambda = \begin{bmatrix} \lambda_{00} & \lambda_{01} & \dots & \lambda_{0s-1} \\ & \lambda_{11} & \dots & \lambda_{1s-1} \\ & & \dots & \dots \\ & & & \lambda_{s-1 s-1} \end{bmatrix}$$

where an element  $\lambda_{ij}$  represents the number of times  $i$ -th level of a factor occurs with the  $j$ -th level of another factor. Solving these equations, together with some more suitably chosen equations (since all the parameters can not be estimated) one would be able to calculate sum of squares due to each main effect and also total sum of squares due to all main effects which on subtraction from total sum of squares will give error sum of squares.

Two examples of partially balanced arrays are given below:

TABLE 18. PARTIALLY BALANCED ARRAYS

(1)						(2)						
factors	assemblies					factors	assemblies					
	1	2	3	4	5		1	2	3	4	5	6
A	0	1	0	1	1	A	0	1	0	1	1	1
B	0	0	1	1	1	B	0	0	1	1	1	1
C	0	1	1	0	1	C	0	1	1	0	1	1
D	0	1	1	1	0	D	0	1	1	1	0	1
						E	0	1	1	1	1	0

In the first array, four factors have been accommodated in 5 assemblies, while in the second 5 factors have been accommodated in 6 assemblies.

I am grateful to Dr. C. R. Rao for suggesting the problems to me and for his guidance.

## REFERENCES

- BOSE, R. C. and BUSH, K. A. (1952): Orthogonal arrays of strength two and three. *Ann. Math. Stat.*, **23**, 508.
- BUSH, K. A. (1952): Orthogonal arrays of index unity. *Ann. Math. Stat.*, **23**, 426.
- DAVIES, O. W. (1954): *Design and Analysis of Industrial Experiments*. Oliver and Boyd, London.
- DAVIES, O. W. and HAY (1950): The construction and uses of fractional factorial designs in industrial research. *Biometrika*, **6**, 233.
- FINNEY, D. J. (1945): The fractional replication of factorial arrangements. *Ann. Eugen.*, **12**, 291.
- NAIR, K. R. and RAO, C. R. (1948): Confounding in asymmetrical factorial experiments. *J. Roy. Stat. Soc., B*, **10**, 109.
- (1941): *Science and Culture*, **7**, 361.
- (1942a): *Science and Culture*, **7**, 457.
- (1942b): *Science and Culture*, **7**, 568.
- PLACKETT and BURMAN (1946): The design of multifactorial experiments. *Biometrika*, **33**, 305.
- RAO, C. R. (1946): On hypercubes of strength  $d$  and a system of confounding in factorial experiments. *Bull. Cal. Math. Soc.*, **38**, 67.
- (1947): Factorial experiments derivable from combinatorial arrangements of arrays. *J. Roy. Stat. Soc., (Suppl.)* **9**, 128.
- (1950): The theory of fractional replication in factorial experiments. *Sankhyā*, **10**, 81.
- (1952): *Advanced Statistical Methods in Biometric Research*. John Wiley and Sons, New York.
- TAGUCHI, G. (1955): An experiment to find an optimum operational standard for anodising aluminium alloy parts. (unpublished report submitted to Indian Statistical Institute)

# A GENERAL CLASS OF QUASIFACTORIAL AND RELATED DESIGNS

By C. RADHAKRISHNA RAO  
Indian Statistical Institute, Calcutta

## 1. INTRODUCTION

A general class of quasifactorial designs was introduced by Nair and Rao (1942) and used in constructing balanced confounded designs for asymmetrical factorial experiments. The full details leading to the construction of asymmetrical designs are reported in another paper by Nair and Rao (1948). The object of this paper is to construct some useful quasifactorial designs for varietal trials. A number of designs closely resembling the quasifactorial system have also been given.

A quasifactorial design is defined as follows. There are  $v = p_1 \times p_2 \times \dots \times p_n$  varieties which can be identified by the multiple system

$$(x_1, x_2, \dots, x_n) \quad \dots \quad (1.1)$$

$$x_i = 1, 2, \dots, p_i; \quad i = 1, \dots, n$$

and  $b$  blocks each containing  $k$  different varieties such that

(i) every variety is used  $r$  times, and

(ii) the two varieties represented by  $(x_1, \dots, x_n)$  and  $(y_1, \dots, y_n)$  occur together in  $\lambda_{c_1, \dots, c_n}$  blocks where  $c_i = 1$  or  $0$  according as  $x_i = y_i$  or  $x_i \neq y_i$ . There are  $(2^n - 1)$  parameters  $\lambda_{c_1, \dots, c_n}$  which need not be all different. The parameters satisfy the following relationships

$$vr = bk \quad \dots \quad (1.2)$$

$$r(k-1) = \sum p(c_1, \dots, c_n) \lambda_{c_1, \dots, c_n}$$

$$p(c_1, \dots, c_n) = (p_{m_1} - 1)(p_{m_2} - 1) \dots (p_{m_s} - 1) \quad \dots \quad (1.3)$$

where  $c_{m_1}, \dots, c_{m_s}$  are unity and the rest zero.

The quasifactorial design as defined above satisfies the parametric requirements of a partially balanced design (Bose and Nair, 1939; Nair and Rao, 1942). We shall consider only the two dimensional quasifactorial designs which are of special interest.



## 2. TWO DIMENSIONAL QUASIFACTORIAL

In the case of two dimensional quasifactorial, the varieties,  $v = p_1 \times p_2$ , can be arranged in a rectangular lattice with  $p_1$  rows and  $p_2$  columns. Given any variety, the rest fall into three groups,  $(p_1-1)$  in the same column,  $(p_2-1)$  in the same row and  $(p_1-1)(p_2-1)$  in the rest of the lattice. These are respectively the first, second and third associates with  $\lambda$  parameters equal to  $\lambda_{01}$ ,  $\lambda_{10}$  and  $\lambda_{11}$ . This is a partially balanced design with the second system of parameters given by

$$p_{ij}^{01} = \begin{pmatrix} p_1-2 & 0 & 0 \\ 0 & 0 & p_2-1 \\ 0 & p_2-1 & (p_2-1)(p_1-2) \end{pmatrix}, p_{ij}^{10} = \begin{pmatrix} 0 & 0 & p_1-1 \\ 0 & p_2-2 & 0 \\ p_1-1 & 0 & (p_1-1)(p_2-2) \end{pmatrix}$$

$$p_{ij}^{11} = \begin{pmatrix} 0 & 1 & p_1-2 \\ 1 & 0 & p_2-2 \\ p_1-2 & p_2-2 & (p_1-2)(p_2-2) \end{pmatrix}.$$

When some of the  $\lambda_{ij}$  are equal, the design may reduce to a partially balanced design with only two associates though not necessarily. Some sufficient conditions for reduction to two associates are given below:

- (i)  $p_1 = p_2, \lambda_{10} = \lambda_{01} = \lambda_1, \lambda_{11} = \lambda_2 \neq \lambda_1$ ,
- (ii)  $\lambda_{10} = \lambda_{11} = \lambda_2, \lambda_{01} = \lambda_1 \neq \lambda_2$  for any  $p_1$  and  $p_2$ ,
- (iii)  $\lambda_{01} = \lambda_{11} = \lambda_2, \lambda_{10} = \lambda_1 \neq \lambda_2$  for any  $p_1$  and  $p_2$ .

The second system of parameters for the case (i) is, using  $p$  for the common value

$$p_{ij}^1 = \begin{pmatrix} p-2 & p-1 \\ p-1 & (p-1)(p-2) \end{pmatrix}, p_{ij}^2 = \begin{pmatrix} 2 & 2(p-2) \\ 2(p-2) & (p-2)^2 \end{pmatrix}$$

and for case (ii)

$$p_{ij}^1 = \begin{pmatrix} p_1-2 & 0 \\ 0 & p_1(p_2-1) \end{pmatrix}, p_{ij}^2 = \begin{pmatrix} 0 & p_1-1 \\ p_1-1 & p_1(p_2-2) \end{pmatrix}$$

and for case (iii).  $p_{ij}$  are obtained by interchanging  $p_1$  and  $p_2$  in the expressions for case (ii). We will consider only designs with three associate classes since most of the partially balanced designs with two associate classes have been listed by Bose, Clatworthy and Shrikhande (1954).

# A GENERAL CLASS OF QUASIFACTORIAL AND RELATED DESIGNS

## 3. CONSTRUCTION OF TWO DIMENSIONAL QUASIFACTORIAL DESIGNS

### 3.1. Series derivable from orthogonal Latin squares.

$$v = pq, b = q(q-1), k = p, r = (q-1) \quad \dots \quad (3.1.1)$$

$$\lambda_{01} = \lambda_{10} = 0, \quad \lambda_{11} = 1.$$

It is known from the work of Bose (1938) and Stevens (1938) that when  $q$  is a prime or a prime power, it is possible to construct  $(q-1)$  orthogonal Latin squares in such a way that they differ only in a cyclical interchange of the rows from the 2nd to the  $q$ -th. Such squares  $(q-1)$  are taken and the rows of each are bordered with numbers  $1, \dots, q$ . In each square there are  $q^2$  cells which may be identified by a pair of integers one representing the row and another the number in the cell (corresponding to the Latin square).

If we represent the varieties by an ordered pair of integers and consider the  $q(q-1)$  columns from all the orthogonal squares as blocks we obtain a design with  $\lambda_{10} = \lambda_{01} = 0$  and  $\lambda_{11} = 1$ . This is because varieties represented by  $(ij)$  and  $(rs)$  occur in no column if  $i = r$  or  $j = s$  and occur in just one column when  $i \neq r$  and  $j \neq s$ . This result can be easily proved by using the special property of orthogonal squares derived by the method of interchanging  $(q-1)$  rows cyclically. As it stands this is a design for  $q^2$  varieties. Omitting  $(q-p)$  rows of the Latin squares designs for  $pq$  varieties with  $\lambda_{10} = \lambda_{01} = 0$  and  $\lambda_{11} = 1$  are obtained. As an illustration, let us consider the designs for  $2 \times 4$ ,  $3 \times 4$  and  $4 \times 4$  obtained from  $4 \times 4$  orthogonal Latin squares.

TABLE 1. DESIGNS FOR  $2 \times 4$ ,  $3 \times 4$  AND  $4 \times 4$

row no.	orthogonal latin squares			
(1)	(2)	(3)	(4)	
1	1 2 3 4	1 2 3 4	1 2 3 4	
2	2 1 4 3	3 4 1 2	4 3 2 1	
3	3 4 1 2	4 3 2 1	2 1 4 3	
4	4 3 2 1	2 1 4 3	3 4 1 2	

The design for  $3 \times 4$  is obtained by omitting the last row and considering the twelve columns. The actual design is

$$(11, 22, 33), (12, 21, 34), (13, 24, 31), (14, 23, 32)$$

$$(11, 23, 34), (12, 24, 33), (13, 21, 32), (14, 22, 31)$$

$$(11, 24, 32), (12, 23, 31), (13, 22, 34), (14, 21, 33)$$

where the 12 varieties are represented by pairs of integers 11, 12, 13, 14, ..., 34.

Another method of construction which is more general than the above is as follows.

Let  $(k-1)$  be the maximum number of orthogonal Latin squares of order  $q$ . Then by superimposing all the squares, with the first row made identical, we obtain a composite square each cell of which contains  $(k-1)$  ordered integers taking values from 1 to  $q$ . Border the columns of such a square with integers from 1 to  $q$  in the same order as they occur in the first row. Omitting the first row we now have, including the bordering elements,  $q(q-1)$  ordered sets of  $k$  elements corresponding to the  $q(q-1)$  cells. To the ordered set corresponding to any cell, we attach integers in the order from 1 to  $k$ , to obtain  $k$  pairs. These  $k$  pairs represent  $k$  varieties of a block. We have  $k \times q$  distinct pairs representing the varieties and  $q(q-1)$  blocks. This provides a quasifactorial design with  $\lambda_{11} = 1$ ,  $\lambda_{10} = \lambda_{01} = 0$ .

As an illustration, let us consider the superimposed two orthogonal squares of order 4 with its columns bordered.

TABLE 2. SUPERIMPOSED ORTHOGONAL SQUARES

1	2	3	4
(1)	(2)	(3)	(4)
11	22	33	44
23	14	41	32
34	43	12	21
42	31	24	13

The design for  $3 \times 4$  with twelve blocks is given in table 2b.

TABLE 2b Quasifactorial design for  $3 \times 4$  with  $\lambda_{11} = 1$ ,  $\lambda_{10} = \lambda_{01} = 0$

order	ordered sets representing blocks											
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
1	1	1	1	2	2	2	3	3	3	4	4	4
2	2	3	4	1	4	3	4	1	2	3	2	1
3	3	4	2	4	3	1	1	2	4	2	1	3

The first block has the varieties (11, 22, 33) the second (11, 23, 34) and so on.

If all the  $(k-1)$  orthogonal squares have a directrix (i.e. have all different elements in the diagonal) then the rows and columns can be bordered with elements in the same order as in the diagonal. Each cell now gives rise to an ordered set of  $(k+1)$  elements, two corresponding to rows and columns and  $(k-1)$  to the orthogonal

# A GENERAL CLASS OF QUASIFACTORIAL AND RELATED DESIGNS

squares. Omitting the diagonal elements and considering the ordered sets in the  $q(q-1)$  remaining cells we can build up as before by attaching integers,  $1, 2, \dots, (k+1)$  for the order a design for  $(k+1) \times q$  varieties in  $q(q-1)$  blocks of  $(k+1)$  plots with  $\lambda_{11} = 1, \lambda_{10} = \lambda_{01} = 0$ .

A design for  $2 \times q$  always exist because it just depends on the existence of a Latin square. Since a Latin square of any order can be written so that it has a directrix, it follows that a design for  $3 \times q$  exists for all  $q$ . Designs for  $4 \times q$  depend on the existence of 3 orthogonal Latin squares of order  $q$  or at least two with a common directrix and so on.

Table 3 below gives the list of useful designs in the series (3.1.1)

$$v = pq, \quad b = q(q-1), \quad k = p, \quad r = (q-1)$$

$$\lambda_{01} = \lambda_{10} = 0, \quad \lambda_{11} = 1.$$

The method of construction adopted (see Table 1) gives the design in groups of blocks representing separate replications wherever such a resolution is possible. The non-resolvable designs which may arise by adopting the method of Table 2 are marked with an asterisk. Designs for  $p = q$  are omitted as they are partially balanced with two classes of associates.

TABLE 3. DESIGNS OF SERIES (3.1.1) ( $p \neq q$ )

sl. no.	$v$	$b$	$k$	$r$	sl. no.	$v$	$b$	$k$	$r$
(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
1	12	12	3	3	11	36	72	4	6
2	15	20	3	4	12	35	42	5	6
3*	18	30	3	5	13	40	56	5	7
4	21	42	3	6	14	45	72	5	8
5	24	56	3	7	15	42	42	6	6
6	27	72	3	8	16	48	56	6	7
7*	30	90	3	9	17	54	72	6	8
8	20	20	4	4	18	56	56	7	7
9	28	42	4	6	19	63	72	7	8
10	32	56	4	7	20	72	72	8	8

We now consider a second series of designs derivable from orthogonal Latin squares with the following parameters

$$v = pq, \quad b = q^2, \quad k = p, \quad r = q \quad \dots \quad (3.1.2)$$

$$\lambda_{11} = 1 = \lambda_{10}, \quad \lambda_{01} = 0.$$



In this case the partially balanced design has only two associate classes as mentioned in section 2. The first and second system of parameters are

$$n_1 = q(p-1), n_2 = (q-1)$$

$$\lambda_1 = 1, \lambda_2 = 0$$

$$p_{ij}^1 = \begin{pmatrix} q(p-2) & q-1 \\ q-1 & 0 \end{pmatrix} \quad p_{ij}^2 = \begin{pmatrix} q(p-1) & 0 \\ 0 & q-2 \end{pmatrix}.$$

It may be seen that this is also a group divisible design with  $p$  groups each containing  $q$  varieties. Two varieties from the same group do not occur in any block while they occur in just one block when they belong to two different groups. We shall not list these designs as the actual plans are given, by Bose, Clatworthy and Shrikhande (1954).

3.2. *Series derivable by the method of joining.* Let us write down the  $p$   $q$  numbers representing varieties in the form of  $p \times q$  rectangular lattice with  $p$  rows and  $q$  columns. Suppose that balanced incomplete block designs exist for the parameters

$$v = p, b = b_1, r = r_1, k, \lambda_1$$

$$v = q, b = b_2, r = r_2, k, \lambda_2$$

then by forming separate designs for varieties in each row and column and combining them we get a quasifactorial design for  $pq$  varieties with parameters

$$b = pb_2 + qb_1, r = r_1 + r_2, k$$

$$\lambda_{11} = 0, \lambda_{01} = \lambda_1, \lambda_{10} = \lambda_2.$$

This design is resolvable if the balanced designs used for each row and column are resolvable. Only special cases are of interest.

$$v = p^2, b = 2p, r = 2, k = p \quad \dots (3.2.1)$$

$$\lambda_{11} = 0, \lambda_{01} = \lambda_{10} = 1.$$

This is Yates' two dimensional square lattice obtained by considering the rows and columns as complete randomised blocks. The partially balanced design has only two associates.

$$v = pq, b = pb_2, k = p, r = 1 + r_2 \quad \dots (3.2.2)$$

$$\lambda_{11} = 0, \lambda_{10} = \lambda_2, \lambda_{01} = 1.$$

In this series the design used for the columns is the randomised block with one replication. The list of useful designs together with the parameters of the balanced design used in the construction is given in Table 4.

## A GENERAL CLASS OF QUASIFACTORIAL AND RELATED DESIGNS

TABLE 4. DESIGNS OF THE SERIES (3.2.2)

sl. no.	parameters of the balanced design				parameters of the quasifactorial design			
	$v=q$	$b_2$	$r_2$	$\lambda_2$	$v=pq$	$b$	$r$	$k=p$
+								
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1	4	4	3	2	12	16	4	3
2	6	10	5	2	18	36	6	3
3	7	7	3	1	21	28	4	3
4	7	14	6	2	21	49	7	3
*5	9	12	4	1	27	45	5	3
6	9	24	8	2	27	81	9	3
7	10	30	9	2	30	100	10	3
8	13	26	6	1	39	91	7	3
9	4	2	2	2	16	12	3	4
10	7	7	4	2	28	35	5	4
11	10	15	6	2	40	70	7	4
12	13	13	4	1	52	65	5	4
*13	16	20	5	1	64	96	6	4
*14	8	14	7	3	32	64	8	4
15	4	4	3	3	16	20	4	4
16	5	5	4	3	20	25	5	4
17	11	11	5	2	55	66	6	5
18	5	3	3	3	25	20	4	5

The designs marked with\* are resolvable.

The designs marked with\* are resolvable.

More general forms of designs obtained by the method of joining are represented by the following system of parameters.

$$r = r_1 + r_2, k \dots (3.2.3)$$

$$v = pq, \quad b = pb_2 + qb_1, \quad r = r_1 + r_2, \quad k$$

$$\lambda_{11} = 0, \quad \lambda_{01} = \lambda_1, \quad \lambda_{10} = \lambda_2.$$

The useful designs of this general class are given in Table 5. All of them are partially balanced with three associate classes except the one corresponding to serial no. 1, which has only two associates.

TABLE 5. DESIGNS OF THE SERIES (3.2.3)

sl. no.	parameters of the balanced design								parameters of the quasifactorial design			
	rows				columns				v	b	r	k
	v = p	b	r <sub>1</sub>	λ <sub>1</sub>	v = q	b <sub>2</sub>	r <sub>2</sub>	λ <sub>2</sub>				
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
1	4	4	3	2	4	4	3	2	16	32	6	3
2	4	4	3	2	6	10	3	1	24	64	8	3
3	4	4	3	2	7	7	3	1	28	56	6	3
4	4	4	3	2	7	14	6	2	28	84	9	3
5	4	4	3	2	9	12	4	1	36	84	7	3
6	4	4	3	2	7	7	4	2	28	42	6	4
7	4	4	2	2	7	7	4	2	40	80	8	4
8	4	4	2	2	10	15	6	2	16	32	5	4
9	4	4	2	2	4	4	3	3	20	40	7	4
10	4	4	2	2	4	5	4	3	25	50	8	4
11	4	4	3	3	5	5	4	3	32	88	9	4
12	5	5	4	3	4	4	2	2				
13	8	14	7	3								

## 4. CIRCULAR LATTICE DESIGNS

4.1. *Construction of designs.* Let us consider  $n$  concentric circles and  $n$  diameters defining  $2n^2$  lattice points on the circles. Each circle has  $2n$  points on it and so also each diameter. If the circles and the diameters are taken as blocks we get a design with the following parameters:

$$v = 2n^2, b = 2n, k = 2n, r = 2.$$

Given any variety the rest fall into three groups, one occurring with it on a circle and on a diameter,  $4(n-1)$  occurring with it either on a circle or on a diameter and  $2(n-1)^2$  not occurring with it. This design, therefore, satisfies the requirements of the first set of parameters of a partially balanced design with

$$n_1 = 1, n_2 = 4(n-1), n_3 = 2(n-1)^2$$

$$\lambda_1 = 2, \lambda_2 = 1, \lambda_3 = 0.$$

It is not difficult to see that requirements of the second system of parameters are also satisfied. The actual matrices are

$$p_{ij}^1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 4(n-1) & 0 \\ 0 & 0 & 2(n-1)^2 \end{pmatrix}, p_{ij}^2 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 2(n-2) & 2(n-1) \\ 0 & 2(n-1) & 2(n-1)(n-2) \end{pmatrix}$$

$$p_{ij}^3 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 4 & 4(n-2) \\ 1 & 4(n-2) & 2(n-2)^2 \end{pmatrix}.$$

There are only four useful designs in this series.

TABLE 6. CIRCULAR LATTICE DESIGNS

sl. no.	$v$	$b$	$k$	$r$
(1)	(2)	(3)	(4)	(5)
1	8	4	4	2
2	18	6	6	2
3	32	8	8	2
4	50	10	10	2

# A GENERAL CLASS OF QUASIFACTORIAL AND RELATED DESIGNS

The method of construction for the design with serial no. 1 is illustrated below by drawing 2 circles and 2 diameters.

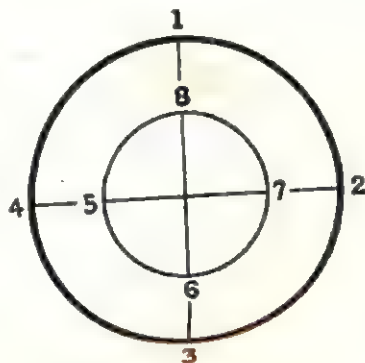


Figure 1.

The blocks are

Circles  
(1, 2, 3, 4)  
(5, 6, 7, 8)

Diameters  
(1, 8, 6, 3)  
(4, 5, 7, 2)

All the designs are resolvable into two separate replications corresponding to the circles and the diameters.

4.2. *Analysis of designs.* Since a circular lattice is partially balanced with three associate classes the general method of analysing such designs could be used. But it is simpler to use the 'P method' explained in Roy and Laha (1956), Rao (1956) as the number of blocks is very small compared to the number of varieties.

Let  $v = 2n^2$ ,  $b = 2n = k$ ,  $r = 2$ .

Denoting the  $n$  circles by  $c_1, \dots, c_n$  and  $n$  diameters by  $d_1, d_2, \dots, d_n$  we define

$B(c_i)$  = the block total corresponding to the circle  $c_i$ .

$B(d_i)$  = the block total corresponding to the diameter  $d_i$ .

$P(c_i) = B(c_i)$ —the sum of mean yields of varieties occurring in  $c_i$ .

$P(d_i)$  = as above for the diameter  $d_i$ .

The estimates of the block constants which need not be computed are

$$b(c_i) = \frac{P(c_i)}{n} - \frac{\sum P(c_i)}{v}$$

$$b(d_i) = \frac{P(d_i)}{n} - \frac{\sum P(d_i)}{v}$$

The sum of squares due to blocks corrected for varieties is

$$\sum b(c_i) B(c_i) + \sum b(d_i) B(d_i)$$

$$= \frac{1}{n} \{ \sum P(c_i) B(c_i) + \sum P(d_i) B(d_i) \} - \frac{1}{v} \{ \sum P(c_i) \sum B(c_i) + \sum P(d_i) \sum B(d_i) \}$$

so that the only quantities need to be computed are the  $B$  and  $P$  values.



The estimate of  $i$ -th varietal effect\* is

$$t_i = \frac{T_i}{2} + \frac{P(c_r) + P(d_s)}{b}$$

where  $c_r$  and  $d_s$  represent the circle and diameter on which the variety  $t_i$  lies.

The variances for comparisons are

$$\begin{aligned} V(t_i - t_j) &= \sigma^2, \text{ if } i, j \text{ are first associates } (\lambda = 2) \\ &= \left(1 + \frac{1}{b}\right) \sigma^2, \text{ if } i, j \text{ are second associates } (\lambda_2 = 1) \\ &= \left(1 + \frac{2}{b}\right) \sigma^2, \text{ if } i, j \text{ are third associates } (\lambda_3 = 0). \end{aligned}$$

The average variance of all comparisons is

$$\frac{v-1}{v+k-3} \sigma^2$$

which may be used to test all varietal differences if the correspondence between the varieties and the integers in the plan of the design is made at random.

It may be observed that the circular lattice designs can also be obtained by considering the dual of a group divisible design with  $2n$  varieties in  $2n^2$  blocks of 2 plots.

$$\begin{aligned} &(i, n+1)(i, n+2) \dots (i, 2n) \\ &(i, n+1)(i, n+2) \dots (i, 2n) \quad i = 1, \dots, n. \end{aligned}$$

This suggests another type of design obtained by daulising the design with the above blocks repeated thrice instead of twice. The parameters of the new design are

$$v = 3n^2, \quad b = 2n, \quad k = 3n, \quad r = 2$$

with three classes of associates  $\lambda_1 = 2, \lambda_2 = 1, \lambda_3 = 0$ . The expressions for the estimates of varietal differences and their variances can be obtained by following the method adopted in the case of circular lattice.

The circular lattice can also be deduced by considering a square lattice and replacing each variety by a pair of varieties. But what is of interest is the simplicity in the analysis of these designs. The diagrammatic representation of the design as a circular lattice provides the association scheme.

#### REFERENCES

- BOSE, R. C. (1938): On the application of the properties of Galois fields to the construction of hyper-Graeco-Latin squares. *Sankhyā*, **3**, 323.
- AND NAIR, K. R. (1939): Partially balanced incomplete block designs. *Sankhyā*, **4**, 337—35.
- BOSE, R. C., CLATWORTHY, W. H. AND SHRIKAND, S. S. (1954): Tables of partially balanced designs with two associate classes. *Tech. Bul. No. 107*, North Carolina Agricultural Experiment Station.
- NAIR, K. R. AND RAO, C. R. (1941a): A general class of quasifactorial designs leading to confounded designs for factorial experiments. *Science and Culture*, **7**, 457.
- (1942a): A note on partially balanced incomplete block designs. **7**, 568.
- (1948): Confounding in asymmetrical factorial experiments. *J. Amer. Stat. Soc.*, **10**, 109.
- RAO, C. R. (1956): On the recovery of inter-block information in varietal trials. *Sankhyā*, **17**, 105-114.
- ROY, J. AND LAHA, R. G. (1956): Classification and analysis of linked block designs. *Sankhyā*, **17**, 115-132.
- STEWENS, W. L. (1938): The completely orthogonalised Latin square. *Ann. Eugen. Lond.*, **9**, 82.

# TWO ASSOCIATE PARTIALLY BALANCED DESIGNS INVOLVING THREE REPLICATIONS

By J. ROY and R. G. LAHA  
Indian Statistical Institute, Calcutta

## 1. SUMMARY

Partially Balanced Incomplete Block (PBIB) designs with two associate classes and involving a few replications are of practical importance. Bose (1951) obtained the complete class of two associate PBIB designs involving two replications. Bose and Clatworthy (1955) recently derived all such designs with  $k > r = 3$  and  $\lambda_1 = 1$  and  $\lambda_2 = 0$ .

In the present paper two associate PBIB designs involving three replications are completely enumerated.

## 2. INTRODUCTION

PBIB designs were introduced by Bose and Nair (1939) and extended by Nair and Rao (1942). Recently for PBIB designs with two associate classes a less demanding definition has been given by Bose and Clatworthy (1955) which is as follows: An arrangement of  $v$  treatments in  $b$  blocks each of  $k$  plots is said to be a PBIB design with two associate classes if

(i) each treatment occurs in  $r$  blocks and no treatment occurs more than once in each block.

(ii) any two treatments are either first or second associates.

(iii) each treatment has  $n_1$  first associates and  $n_2$  second associates.

(iv) for any pair of treatments which are  $i$ -th associates the number  $p_{11}^i$  of treatments which are simultaneously the first associate of both the treatments is independent of the pair of treatments with which we start;  $i = 1, 2$ .

(v) any pair of treatments which are  $i$ -th associates occur together in exactly  $\lambda_i$  blocks;  $i = 1, 2$ .

In such a case it has been shown that the number  $p_{jk}^i$  of treatments which are simultaneously the  $j$ -th associates of a treatment  $\theta$  and  $k$ -th associates of a treatment  $\phi$  where  $\theta$  and  $\phi$  are themselves  $i$ -th associates is independent of  $\theta$  and  $\phi$ . These parameters satisfy the following conditions:

$$\left. \begin{aligned} vr &= bk \\ v &= n_1 + n_2 + 1 \\ \lambda_1 n_1 + \lambda_2 n_2 &= r(k-1) \end{aligned} \right\} \dots (2.1)$$

$$\left. \begin{aligned} p_{11}^1 + p_{12}^1 + 1 &= p_{11}^2 + p_{12}^2 = n_1 \\ p_{12}^1 + p_{22}^1 &= p_{12}^2 + p_{22}^2 + 1 = n_2 \\ n_1 p_{12}^1 &= n_2 p_{11}^2 \\ n_1 p_{22}^1 &= n_2 p_{12}^2 \end{aligned} \right\} \dots (2.2)$$

To avoid triviality, we shall take  $n_1, n_2 > 0$  and  $\lambda_1 \neq \lambda_2$  or, without loss of generality,  $\lambda_1 > \lambda_2$ .

For a PBIB design with two associate classes, let the 'incidence matrix' be denoted by  $N = (n_{ij})$  where  $n_{ij} = 1$  or 0 according as the  $j$ -th treatment occurs in the  $i$ -th block or not  $i = 1, 2, \dots, b; j = 1, 2, \dots, v$ . It has been shown by Connor and Clatworthy (1954) that

$$|N'N| = rk(r-x_1)^{\alpha_1}(r-x_2)^{\alpha_2} \quad \dots (2.3)$$

where

$$x_i = \frac{1}{2} [(\lambda_1 + \lambda_2) + (\lambda_1 - \lambda_2)\{-\gamma + (-1)^i \sqrt{\Delta}\}] \quad \dots (2.4)$$

$$\alpha_i = [(v-1)\{(-1)^i \gamma + \sqrt{\Delta} + 1\} - 2n_i] / 2\sqrt{\Delta} \quad \dots (2.5)$$

$(i = 1, 2)$

where

$$\left. \begin{aligned} \gamma &= p_{12}^2 - p_{12}^1 \\ \beta &= p_{12}^2 + p_{12}^1 \\ \Delta &= \gamma^2 + 2\beta + 1. \end{aligned} \right\} \quad \dots (2.6)$$

The numbers  $\alpha_1, \alpha_2$  must be positive integers and in general

$$r > x_2 > x_1 \quad \dots (2.7)$$

but if  $b < v$

$$r = x_2 \quad \dots (2.8)$$

and

$$b > v - \alpha_2 = \alpha_1 + 1. \quad \dots (2.9)$$

For designs with  $b < v$  the relation  $r = x_2$  may be written alternatively as

$$(\lambda_1 - \lambda_2)\{(r - \lambda_2)p_{12}^1 - (r - \lambda_1)p_{12}^2\} = (r - \lambda_1)(r - \lambda_2) \quad \dots (2.10)$$

a result first obtained by Nair (1943). In general, we have however

$$(r - \lambda_1)(r - \lambda_2) > (\lambda_1 - \lambda_2)\{(r - \lambda_2)p_{12}^1 - (r - \lambda_1)p_{12}^2\} \quad \dots (2.11)$$

A two associate PBIB design will be said to be 'connected' if the matrix  $C = rI - \frac{1}{k} N'N$  is of rank  $(v-1)$ . Since the latent roots of the matrix  $C$  are 0 and  $r(1-1/k) + x_i/k$  of multiplicity  $\alpha_i$  ( $i = 1, 2$ ) it follows that a necessary and sufficient condition for connectivity is

$$r(k-1) + x_i > 0 \text{ for } i = 1, 2 \quad \dots (2.12)$$

## TWO ASSOCIATE PBIB DESIGNS WITH THREE REPLICATIONS

### 3. THREE REPLICATE TWO ASSOCIATE PBIB DESIGNS WITH $k > 3$

For three replicate two associate PBIB designs the parameters  $(\lambda_1, \lambda_2)$  can take only the following different pairs of values: (i) (3, 2) (ii) (3, 1) (iii) (3, 0) (iv) (2, 1) (v) (2, 0) and (vi) (1, 0). Of these six types (iii) is not connected and (i) and (ii) can be derived from corresponding Balanced Incomplete Block (BIB) designs by replacing each treatment by a group of treatments. This is discussed in sub-section 3.1. If we consider designs with  $k > 3$ , then of the remaining possibilities (iv), (v) and (vi), the type (vi) has been completely enumerated by Bose and Clatworthy (1955). There it is shown that:—*Three replicate two associate PBIB designs with  $k > 3$  and  $\lambda_1 = 1$ ,  $\lambda_2 = 0$  must belong to one of the following classes:*

- (a) *Designs obtained by dualising BIB designs with parameters  $k^* = 3$  and  $\lambda^* = 1$*
- (b) *Lattice designs with three replications*
- (c) *The design with parameters*

$$v = 45, b = 27, r = 3, k = 5, n_1 = 12, n_2 = 32$$

$$\begin{bmatrix} p_{11}^1 = 3 & p_{12}^1 = 18 \\ & p_{22}^1 = 24 \end{bmatrix} \quad \begin{bmatrix} p_{11}^2 = 3 & p_{12}^2 = 9 \\ & p_{22}^2 = 22 \end{bmatrix}.$$

We shall show in sub-section 3.3 that the case (v) is impossible and obtain in sub-section 3.2 all designs of the type (iv).

3.1. *Designs with  $\lambda_1 = 3$ .* It is well-known that if in a BIB design  $D^*$  with parameters

$$v^* = m, b^* = b, r^* = 3, k^* = p, \lambda^* = \lambda (\lambda = 1, 2)$$

each treatment is replaced by a group of  $n$  treatments the resulting design  $D$  is a two associate PBIB with the following parameters

$$\begin{aligned} v &= mn, b = b, r = 3, k = pn \\ \lambda_1 &= 3, \lambda_2 = \lambda \\ n_1 &= n-1, n_2 = n(m-1), p_{12}^1 = 0 \end{aligned}$$

Conversely, it can be easily verified that any two associate PBIB design with  $r = 3$  and  $\lambda_1 = 3, \lambda_2 = \lambda (\lambda = 1, 2)$  can always be obtained in like manner from a BIB design of the type  $D^*$ . It is also easy to check that any two associate PBIB design with  $r = 3, \lambda_1 = 3$  and  $\lambda_2 = 0$  is disconnected.

It is also known that there are only three BIB designs of the type  $D^*$ . Namely

$$v^* = 4, b^* = 6, k^* = 2, r^* = 3, \lambda^* = 1 \quad \dots \quad (3.1)$$

$$v^* = b^* = 7, k^* = r^* = 3, \lambda^* = 1 \quad \dots \quad (3.2)$$

$$\text{and} \quad v^* = b^* = 4, k^* = r^* = 3, \lambda^* = 2 \quad \dots \quad (3.3)$$



Therefore we have the following:

**Theorem 3.1.** *All two associate PBIB designs with three replications and  $\lambda_1 = 3$  are derivable from the BIB designs (3.1), (3.2) or (3.3) by replacing each treatment by a group of  $n$  treatments.*

**3.2.** *Designs with  $\lambda_1 = 2$  and  $\lambda_2 = 1$ .*

From (2.1) we get  $bk = 3v = 3(n_1 + n_2 + 1)$

$$3(k-1) = 2n_1 + n_2.$$

Eliminating  $n_2$  we have

$$b = 9 - t \quad \dots (3.4)$$

$$v = 3k - kt/3 \quad \dots (3.5)$$

where  $t$  is given by

$$n_1 = kt/3 - 2. \quad \dots (3.6)$$

Since  $k > 3$ , that is  $b < v$ , we have from (2.10)

$$2p_{12}^1 - p_{12}^2 = 2. \quad \dots (3.7)$$

Using the relations

$$p_{12}^1 = n_2 p_{11}^2 / n_1 \text{ and } p_{12}^2 = n_1 - p_{11}^2$$

and substituting in (3.7) we get

$$p_{11}^2 = \frac{t(kt-6)}{9(6-t)}. \quad \dots (3.8)$$

From (2.4) and (2.8) we get :

$$3 = r = x_2 = \frac{1}{2}(3 - \gamma + \sqrt{\Delta})$$

so that

$$\sqrt{\Delta} = 3 + \gamma.$$

Substituting in (2.5), we get

$$\alpha_1 = \frac{n_1 + 2n_2}{3 + \gamma}.$$

But

$$\begin{aligned} 3 + \gamma &= 3 + p_{12}^2 - p_{12}^1 = p_{12}^1 + 1 \text{ because of (3.7)} \\ &= n_2 p_{11}^2 / n_1 + 1 \end{aligned}$$

# TWO ASSOCIATE PBIB DESIGNS WITH THREE REPLICATIONS

and substituting the values of  $n_1, n_2$  and  $p_{11}^2$ ,

$$\text{we get } \alpha_1 = \frac{9k(6-t)^2}{(kt+6)(9-t)-kt^2} \quad \dots (3.9)$$

Let us write

$$kt/3 = u, \text{ or, } k = 3u/t. \quad \dots (3.10)$$

Since  $n_1$  is a positive integer, from (3.6), we find that  $u$  must be a positive integer greater than 2. From (3.8) we get, in terms of  $u$

$$p_{11}^2 = \frac{t(u-2)}{3(6-t)} = \frac{2-u}{3} + \frac{2(u-2)}{6-t} \quad \dots (3.11)$$

Since  $3p_{11}^2$  is a positive integer, so must be

$$\frac{6(u-2)}{6-t} = w, \text{ (say.)}$$

$$\text{Then } u = w + 2 - \frac{tw}{6} \quad \dots (3.12)$$

so that

$$p_{11}^2 = \frac{wt}{18}. \quad \dots (3.13)$$

We also find from (3.8) that the only admissible values of  $t$  are 1, 2, 3, 4 and 5. For each such value of  $t$ ,  $w$  must make  $wt/18$  positive integral, and the corresponding value of  $u$  given by (3.12) should make

$$\alpha_1 = \frac{3k(6-t)^2}{(u+2)(9-t)-ut} \quad \dots (3.14)$$

a positive integer which because of (2.9) must not be greater than  $b-1$ .

Let us take up the different possible values of  $t$  one by one. Consider the case  $t = 1$ . In this case  $p_{11}^2 = w/18 = a$  (say), so that  $w = 18a$ . Therefore we get  $u = 15a + 2$  and  $k = 3u = 3(15a + 2)$  so that

$$\alpha_1 = \frac{15(15a+2)}{7a+2} = 7 + \frac{16(11a+1)}{7a+2}.$$

But here  $b = 8$  and  $\alpha_1 > b-1$  for all positive values of  $a$ . Hence no design is possible.

For the case  $t = 2$  we have  $p_{11}^2 = w/9 = a$ , say. Then  $w = 9a$ ,  $u = 6a + 2$ ,  $k = 3u/2 = 3(3a + 1)$  so that

$$\alpha_1 = \frac{24(3a+1)}{5a+4} = 6 + \frac{42a}{5a+4}.$$

But here  $b = 7$  and  $\alpha_1 > b - 1$  for all positive values of  $a$ . Hence in this case also no design is available.

Now take the case  $t = 3$ . Here  $p_{11}^2 = w/6 = a$  say. Then  $w = 6a$ ,  $u = 3a + 2$ ,  $k = u = 3a + 2$  so that

$$\alpha_1 = \frac{3(3a+2)}{a+2}.$$

Here  $b = 6$  and since

$$b - 1 - \alpha_1 = \frac{4(1-a)}{a+2}$$

has to be non-negative, the only permissible value of  $a$  is  $a = 1$ . Fortunately this value of  $a$  makes  $\alpha_1 = 5$  a positive integer. For this value of  $a$ , we have  $k = 5$ ,  $v = 10$ ,  $n_1 = 3$  and  $p_{11}^2 = 1$ . We have merely proved that the above parameters satisfy all the known necessary restrictions for the existence of the corresponding design. This design however does exist and is listed in Bose, Clatworthy and Shrikhande (1954) as Design number T9.

For  $t = 4$ ,  $p_{11}^2 = 2w/9$  which must therefore be an even integer  $= 2a$ , say. Then  $w = 9a$ ,  $u = 3a + 2$  and  $k = 3u/4 = (9a + 6)/4 = 2a + 1 + (a + 2)/4$ . Since this is integral  $a$  itself must be of the form  $4c + 2$ . Thus  $p_{11}^2 = 4(2c + 1)$ ,  $w = 9(4c + 2)$ ,  $u = 4(3c + 2)$  and  $k = 3(3c + 2)$ . Thus

$$\alpha_1 = \frac{6(3c+2)}{2c+3}.$$

Here  $b = 5$  and therefore  $b - 1 - \alpha_1 = \frac{-10c}{2c+3}$  which is non-negative only when  $c = 0$ . If  $c = 0$ ,  $\alpha_1 = 4$  is integral. For this case, we get  $k = 6$ ,  $v = 10$ ,  $n_1 = 6$  and  $p_{11}^2 = 4$ . This design is known to exist and listed as Design number T15 in Bose, Clatworthy and Shrikhande (1954).

When  $t = 5$ , we get  $p_{11}^2 = 5w/18$  which must be divisible by 5,  $p_{11}^2 = 5a$  say. Then  $w = 18a$ ,  $u = 3a + 2$  and  $k = 3u/5 = (9a + 6)/5 = 2a + 1 - (a - 1)/5$ . Hence  $a$  must be of the form  $a = 5c + 1$  so that  $k$  may be integral. Then  $p_{11}^2 = 5(5c + 1)$ ,  $w = 18(5c + 1)$ ,  $u = 5(3c + 1)$  and  $k = 3(3c + 1)$  so that

$$\alpha_1 = \frac{3(3c+1)}{1-5c}$$

## TWO ASSOCIATE PBIB DESIGNS WITH THREE REPLICATIONS

The only value of  $c$  for which this is positive integral is  $c = 0$ . But if  $c = 0$ ,  $k = 3$  which contradicts our assumption that  $k > 3$ . Hence no such design exists. We may summarise the results of this section in the following.

**Theorem 3.2.** *There are only two PBIB designs with two associate classes and  $r = 3$ ,  $\lambda_1 = 2$ ,  $\lambda_2 = 1$  and  $k > 3$ , namely:*

$$T_9 : v = 10, k = 5, n_1 = 3, p_{11}^2 = 1$$

and

$$T_{15} : v = 10, k = 6, n_1 = 6, p_{11}^2 = 4.$$

**3.3. Impossibility of designs with  $\lambda_1 = 2$  and  $\lambda_2 = 0$ .** Suppose that such a design exists. Then from (2.10) we get

$$2(3p_{12}^1 - p_{12}^2) = 3.$$

But since  $p_{12}^1$  and  $p_{12}^2$  are non-negative integers, clearly such a thing is impossible. Therefore, we get the

**Theorem 3.3.** *No PBIB design with two associate classes and  $r = 3$ ,  $\lambda_1 = 2$ ,  $\lambda_2 = 0$  and  $k > 3$  exists.*

### 4. THREE REPLICATE TWO ASSOCIATE PBIB DESIGNS WITH $k = 3$

**4.1. The case  $\lambda_1 = 2, \lambda_2 = 1$ .** Here, we get from (2.1)  $2n_1 + n_2 = 6$  and therefore there are only two possibilities namely (i)  $n_1 = n_2 = 2$  and (ii)  $n_1 = 1, n_2 = 4$ .

In case (i), if we set  $p_{12}^1 = t$ , we get from (2.2)  $p_{11}^1 = 1 - t$  and  $p_{22}^2 = t - 1$ . Hence the only permissible value of  $t$  is  $t = 1$ . Then we have  $v = b = 5$ ,  $p_{12}^1 = 1$  and  $\alpha_1 = 2$  which is integral.

Writing the digits 1, 2, ..., 5 for the treatments, a plan for the design (4.1) obtained by taking the following triplets as blocks:

$$(1, 2, 3); (1, 2, 4); (1, 3, 5); (2, 4, 5); (3, 4, 5) \quad \dots \quad (4.1)$$

In case (ii) we have from (2.2)

$$p_{11}^1 + p_{12}^1 = 0$$

Therefore

$$p_{11}^1 = p_{12}^1 = p_{11}^2 = 0,$$

and we get  $v = b = 6$  and  $\alpha_1 = 2$  is integral. A plan of the design is

$$(1, 2, 3); (1, 2, 4); (1, 5, 6); (2, 5, 6); (3, 4, 5); (3, 4, 6) \quad \dots \quad (4.2)$$

where the triplets form the blocks.

We may summarise these results in the following



Theorem 4.1. *There are only two two-associate PBIB designs with  $r = k = 3$  and  $\lambda_1 = 2, \lambda_2 = 1$  namely*

$$v = b = 5, n_1 = n_2 = 2, p_{12}^1 = 1$$

and  $v = b = 6, n_1 = 1, n_2 = 4, p_{12}^1 = 0$

with plans given in (4.1) and (4.2) respectively.

4.2. The case  $\lambda_1 = 2, \lambda_2 = 0$ . Here we get  $n_1 = 3$  and from (2.11)  $2(3p_{12}^1 - p_{12}^2) \leq 3$  so that the permissible values are  $3p_{12}^1 - p_{12}^2 \leq 1$ . Setting  $p_{12}^1 = t$ , we get from (2.2)  $p_{12}^2 = 3 - \frac{3t}{n_2}$ . Hence  $3t(1 + 1/n_2) \leq 4$  and the only possibility is  $t = 1$  and  $n_2 = 3$ . This gives  $v = b = 7$  and  $\alpha_1 = 3 - 3/2\sqrt{2}$  which is not integral. Hence, the

Theorem 4.2. *There is no two associate PBIB design with  $r = k = 3$  and  $\lambda_1 = 2, \lambda_2 = 0$ .*

4.3. The case  $\lambda_1 = 1, \lambda_2 = 0$ . Here  $n_1 = 6$  and from (2.11) we get  $3p_{12}^1 - 2p_{12}^2 \leq 6$ . Setting  $p_{12}^1 = t$ , we get from (2.2)  $p_{12}^2 = 6 - \frac{6t}{n_2}$  and therefore the following restriction on  $t$  and  $n_2$ .

$$t(1 + 4/n_2) \leq 6 \quad \dots (4.3)$$

Since  $p_{11}^1 = 5 - t$ , the permissible values of  $t$  are 0, 1, 2, 3, 4 and 5. Let us take up these values of  $t$  one by one.

If  $t = 0$ , write  $n_2 = a$ . Then we get  $v = 7 + a, \gamma = 6, \beta = 6$  and  $\sqrt{\Delta} = 7$  so that  $\alpha_1 = a/7$  on simplification. Since this is integral  $a$  must be a multiple of 7,  $a = 7(c-1)$  say. Then  $v = b = 7c$  and the values of  $x_1, x_2$  defined in (2.4) come out to be  $-6$  and  $+1$  respectively.

But

$$r = k = 3$$

and therefore

$$r(k-1) + x_1 = 6 + (-6) = 0$$

which contradicts the condition (2.12) for connectivity. Therefore there is no connected design in this case.

If  $t = 1$ , since 6 must be divisible by  $n_2$ , the permissible values of  $n_2$  are 1, 2, 3 and 6. The only value of  $n_2$  that makes  $\alpha_1$  integral is  $n_2 = 1$  which gives  $v = b = 8, p_{12}^1 = 1, p_{12}^2 = 0$ . This design exists and is listed as Design number R5 in Bose, Clatworthy and Shrikhande (1954).

If  $t = 2$ , we get from (4.3)  $4/n_2 \leq 2$  and since  $6t = 12$  must be divisible by  $n_2$  the only permissible value are  $n_2 = 2, 3, 4, 6$ , and 12. Amongst these, the values of  $n_2$  that make  $\alpha_1$  integral are  $n_2 = 2$  and  $n_2 = 3$  respectively. If  $n_2 = 2$ , we get

## TWO ASSOCIATE PBIB DESIGNS WITH THREE REPLICATIONS

$v = b = 9$ ,  $p_{12}^2 = 0$ ; this is a triple lattice design. If  $n_2 = 3$ , we get  $v = b = 10$  and this is listed as Design number T6 in Bose, Clatworthy and Shrikhande (1954).

When  $t = 3$  we get from (4.3)  $4/n_2 \leq 1$ . Since  $6t = 18$  must be divisible by  $n_2$ , permissible values of  $n_2$  are  $n_2 = 6, 9$  and  $18$ . For  $n_2 = 6$  and  $9$ ,  $\alpha_1$  is integral. When  $n_2 = 6$ , we get  $v = b = 13$ ,  $p_{12}^2 = 3$ ; this has the Design number C1 in Bose, Clatworthy and Shrikhande (1954). When  $n_2 = 9$  we get,  $v = b = 16$  and  $p_{12}^2 = 4$ . This design exists and is listed as Design number LS14 in Bose, Clatworthy and Shrikhande (1954).

When  $t = 4$ , we must have  $4/n_2 \leq 1/2$  and  $n_2$  must be a factor of  $24$ . Thus  $n_2 = 8, 12$  and  $24$ ; Thus  $n_2 = 8, 12$  and  $24$ ; of these the only value that makes  $\alpha_1$  integral is  $n_2 = 8$ . When  $n_2 = 8$ , we get  $v = b = 15$ ,  $p_{12}^2 = 3$ . This is Design number T28 in Bose, Clatworthy and Shrikhande (1954).

In the case  $t = 5$ ,  $4/n_2 \leq 1/5$  and  $n_2$  must be a factor of  $30$ . Thus  $n_2 = 30$  but this value does not make  $\alpha_1$  integral. Therefore there is no design in this class.

We may summarise our results in the following

**Theorem 4.3.** *The class of two-associate PBIB design with  $r = k = 3$  and  $\lambda_1 = 1, \lambda_2 = 0$  contains only the designs numbered R5, T6, C1, LS 14 and T28 in Bose, Clatworthy and Shrikhande (1954) and the triple lattice design with  $v = b = 9$ .*

It is interesting to note that Bose, Clatworthy and Shrikhande (1954) give another design, namely Design number S11 with parameters  $v = b = 19$ ,  $r = k = 3$ ,  $\lambda_1 = 1, \lambda_2 = 0, n_1 = 6, n_2 = 12$

$$\begin{bmatrix} p_{11}^1 = 1 & p_{12}^1 = 14 \\ & p_{22}^1 = 8 \end{bmatrix} \quad \begin{bmatrix} p_{11}^2 = 2 & p_{12}^2 = 4 \\ & p_{22}^2 = 7 \end{bmatrix}$$

and two other designs numbered S12 and S13 are derived from it by respectively duplicating and triplicating this design. This is not covered by our theorem. That a two associate PBIB design with these parameters is impossible can be easily demonstrated by calculating the value of  $\alpha_1$  which turns out to be  $\alpha_1 = 9 + 3/\sqrt{17}$  which is not integral.

### 5. THREE REPLICATE TWO ASSOCIATE PBIB DESIGNS WITH $k = 2$

5.1. *The case  $\lambda_1 = 2, \lambda_2 = 1$ .* In this case we get from (2.1) the relation  $2n_1 + n_2 = 3$ , so that the only solution is  $n_1 = n_2 = 1$  and therefore  $v = 3$ . But  $b = rv/k = 9/2$  becomes fractional. Hence no such design is possible.

5.2. *The case  $\lambda_1 = 2, \lambda_2 = 0$ .* Such designs are clearly impossible since (2.1) leads to  $n_1 = 3/2$ .

5.3. *The case  $\lambda_1 = 1, \lambda_2 = 0$ .* In this case we get from (2.1)  $n_1 = 3$ . Setting  $p_{12}^1 = t$ , we get from (2.2)  $p_{11}^1 = 2 - t$  and  $p_{12}^2 = 3 - 3t/n_2$ . Therefore the permissible values of  $t$  are 0, 1 and 2. But from (2.11) we get on simplification

$$t(1 + 2/n_2) \leq 4. \quad \dots \dots \dots (5.1)$$

Thus if  $t=0$  we get on simplification  $\alpha_1 = \frac{v}{4} - 1$  and therefore  $v$  must be a multiple of 4,  $v = 4a$ , say. Then  $n_2 = 4(a-1)$  and the values of  $x_1, x_2$  defined in (2.4) turn out to be  $-3$  and  $1$  respectively. But this makes

$$r(k-1) + x_1 = 3 + (-3) = 0$$

violating the condition (2.12) of connectivity. Hence no connected design is possible with  $t=0$ . If  $t=1$ , since  $n_2$  divides  $3t$ ,  $n_2 = 1, 3$ , but none of these makes  $\alpha_1$  an integer. If  $t=2$ ,  $n_2$  must divide 6 and at the same time satisfy (5.1). Thus  $n_2 = 2, 3$  and  $6$ . Of these values only  $n_2 = 2$  and  $n_2 = 6$  make  $\alpha_1$  integral. If  $n_2 = 2$  we get  $v = 6, b = 9, p_{12}^2 = 0$ . This design is immediately recognised to be the dual of the simple  $3^2$  lattice design and the blocks are given by the pairs

$$(1, 4), (1, 5), (1, 6), (2, 4), (2, 5), (2, 6), (3, 4), (3, 5), (3, 6) \dots \quad (5.2)$$

where the integers 1, 2, ..., 6 denote the six treatments.

If  $n_2 = 6$ , we get  $v = 10, b = 15, p_{12}^2 = 2$ . A plan for such a design is given below:

$$\left. \begin{array}{l} (1, 8), (2, 6), (3, 5), (4, 5), (5, 10) \\ (1, 9), (2, 7), (3, 7), (4, 6), (6, 9) \\ (1, 10), (2, 10), (3, 9), (4, 8), (7, 8) \end{array} \right\} \dots \quad (5.3)$$

where the treatments are indicated by the integers 1, 2, ..., 10 and the pairs are the blocks.

We therefore have the following

**Theorem 5.1.** *There are only two two-associate PBIB designs with  $r = 3$  and  $k = 2$  namely*

$$v = 6, b = 9, \lambda_1 = 1, \lambda_2 = 0, n_1 = 3, n_2 = 2, p_{12}^2 = 0$$

$$\text{and } v = 10, b = 15, \lambda_1 = 1, \lambda_2 = 0, n_1 = 3, n_2 = 6, p_{12}^2 = 2$$

*with plans given by (5.2) and (5.3) respectively.*

#### REFERENCES

- BOSE, R. C. (1951): Partially balanced incomplete block designs with two associate classes and involving only two complete replications. *Bull. Cal. Stat. Ass.*, **3**, 120-125.
- AND CLATWORTHY, W. H. (1955): Some classes of partially balanced designs. *Ann. Math. Stat.*, **26**, 212-231.
- AND SHRIKHANDE, S. S. (1954): Tables of partially balanced designs with two associate classes. *Tech. Bull.* No. 107 (1954) North Carolina Agricultural Experiment Station.
- BOSE, R. C. AND NAIR, K. R. (1939): Partially balanced incomplete block designs. *Sankhyā*, **4**, 337-372.
- CONNOR, W. S. AND CLATWORTHY, W. H. (1954): Some theorems for partially balanced designs. *Ann. Math. Stat.*, **25**, 100-112.
- NAIR, K. R. (1943): Certain inequality relationships among the combinatorial parameters of incomplete block design. *Sankhyā*, **6**, 255-259.
- AND RAO, C. R. (1942): A note on partially balanced incomplete block designs. *Science and Culture*, **7**, 568-569.

*Paper received : February, 1956.*



# SOME SERIES OF BALANCED INCOMPLETE BLOCK DESIGNS

By D. A. SPROTT

Toronto, Canada

## 1. INTRODUCTION

A balanced incomplete block (BIB) design is defined as an arrangement of  $v$  objects in  $b$  blocks of  $k$  distinct objects each, so that there are  $r$  blocks containing any given object and  $\lambda$  blocks containing any two given objects. Such designs have been studied for their applications to statistics, where the objects are usually varieties, and for their combinatorial use.

Various methods of construction have been studied by Bose (1939, 1942), who developed two module theorems and applied them to form several one-parameter families of designs. Sprott (1954) used the first module theorem to derive some two-parameter families of designs which included as special cases many of Bose's original series. It is the purpose of this paper to use the methods of Sprott (1954) and the two module theorems to develop some further series of BIB's. For completeness, Bose's two module theorems are stated in the form in which they will be used.

**First Module Theorem:** Let  $M$  be an additive abelian group of  $n$  elements  $y_u$  and to every element let there correspond one variety  $y_u$ . Suppose that there exist  $m$  initial blocks  $B_1, B_2, \dots, B_m$  such that

- (1) every block contains exactly  $k$  different varieties;
- (2) the differences arising from the  $m$  blocks are symmetrically repeated, each occurring  $\lambda$  times;

then one can form blocks  $B_{j\theta}$  of varieties  $y_w$  where

$$y_w = y_u + \theta, \quad y_u \in B_j, \quad \theta \in M.$$

The set of  $mn$  blocks  $B_{j\theta}$  ( $j = 1, 2, \dots, m, \theta = y_1, y_2, \dots, y_n$ ) forms a BIB with parameters  $v = n, b = mn, r = mk, k, \lambda$ .

**Second Module Theorem:** Let  $M$  be an additive abelian group of  $n$  elements  $y_1, y_2, \dots, y_n$ , and to each  $y_i$  let there correspond one variety. Let there be adjoined a new variety  $\infty$ . Suppose that there exist  $t+s$  initial blocks  $B_1, \dots, B_t, B'_1, \dots, B'_s$  such that

- (1) each  $B_i$  contains exactly  $k$  distinct varieties and each  $B'_i$  contains  $\infty$  and  $k-1$  other distinct varieties;
- (2)  $kt = ns - \lambda$  and  $(k-1)s = \lambda$ ;
- (3) the differences arising from the  $(s+t)$  blocks  $B_1, \dots, B_t, B'_1, \dots, B'_s$  (where  $B''_p = B'_p$  with  $\infty$  deleted) are symmetrically repeated, each occurring  $\lambda$  times;



then the blocks  $B_{j\theta}$  and  $B'_{j\theta}$  can be formed from varieties

$$y_u = y_v + \theta, \quad y_u \in B_j \text{ or } B'_j, \quad \theta \in M, \quad \theta + \infty = \infty.$$

These  $n(s+t)$  blocks  $B_{j\theta}$  and  $B'_{j\theta}$  form BIB with parameters

$$v = n+1, \quad b = n(s+t), \quad r = ns, \quad k, \quad \lambda.$$

## 2. SERIES 1

Theorem 2.1: If  $m(\lambda-1)+1 = p^a$ , where  $p$  is a prime, then the design with parameters

$$v = m(\lambda-1)+1, \quad b = mv, \quad r = mk, \quad k = \lambda \cdot \lambda,$$

can be constructed from the initial blocks

$$(0, x^i, x^{i+m}, x^{i+2m}, \dots, x^{i+m(\lambda-2)})$$

where  $x$  is a primitive element of  $GF(v)$  and  $i$  ranges from 0 to  $m-1$ .

Proof: The differences not involving the zero element are

$$x^{i+(r+s)m} - x^{i+rm} = x^{i+rm}(x^{sm} - 1) = x^{q_s+i+rm}$$

where  $r = 0, 1, 2, \dots, \lambda-2$ ;  $s = 1, 2, \dots, \lambda-2$ .

These differences for  $s$  fixed are all distinct, for if not, there exist  $i, i', r, r'$ , such that  $i = i', r = r'$  do not hold simultaneously, and

$$i+rm \equiv i'+r'm \pmod{m(\lambda-1)}$$

$$i-i' \equiv m(r-r') \pmod{m(\lambda-1)}$$

$$i-i' \equiv 0 \pmod{m}.$$

Hence  $i = i'$  since  $i$  and  $i'$  are less than  $m$ ; therefore

$$r-r' \equiv 0 \pmod{\lambda-1},$$

and therefore  $r = r'$  since  $r$  and  $r'$  are less than  $\lambda-1$ .

The number of differences for  $s$  fixed is  $m(\lambda-1)$ , all distinct; hence the differences for  $s$  fixed range once over  $GF(v)$ , and as  $s$  ranges, the differences are symmetrically repeated, each occurring  $(\lambda-2)$  times. The differences involving the zero element are

$$\pm x^i, \pm x^{i+m}, \pm x^{i+2m}, \dots, \pm x^{i+m(\lambda-2)}.$$

These are  $2m(\lambda-1)$  in number, and since

$$x^{i+rm} \neq x^{i'+r'm},$$

the differences cover  $GF(v)$  twice.

Therefore the differences occur  $\lambda$  times in all, and the design can be constructed by the first module theorem.

# SOME SERIES OF BALANCED INCOMPLETE BLOCK DESIGNS

## 3. SERIES 2

Theorem 3.1: If  $v = 2am(2a\lambda + 1) + 1 = p^n$ , where  $p$  is a prime and if among the  $a\lambda$  exponents  $q_s$ , where

$$x^{2ams} - 1 = x^{q_s} \quad (s = 1, 2, \dots, a\lambda),$$

each residue class (mod  $a$ ) is represented  $\lambda$  times, then the design with parameters

$$v = 2am(2a\lambda + 1) + 1, \quad b = mv, \quad r = m(2a\lambda + 1), \quad k = 2a\lambda + 1, \quad \lambda,$$

can be constructed from the initial blocks

$$(x^{ai}, x^{ai+4am}, \dots, x^{ai+4a^2m\lambda})$$

where  $x$  is a primitive element of  $GF(v)$  and  $i$  ranges from 0 to  $m-1$ .

*Proof:* The differences are expressible in the form

$$x^{ai+2arm+q_s}$$

and

$$x^{q_c+ai+2arm}$$

where

$$c = 2a\lambda + 1 - s$$

and

$$x^{q_s} = x^{2ams} - 1 \quad (s = 1, 2, \dots, a\lambda; \quad r = 0, 1, \dots, a\lambda).$$

Since  $x$  is a primitive element,

$$x^{2am(2a+1)} = 1, \quad x^{am(2a\lambda+1)} = -1,$$

$$x^{2am(2a-1)} + x^{2am(2a-3)} + \dots + x^{2am} + 1 = 0.$$

Hence

$$\begin{aligned} x^{q_s} &= x^{2ams} - 1 \\ &= (x^{2am} - 1)(x^{2am(s-1)} + x^{2am(s-2)} + \dots + x^{2am} + 1). \end{aligned}$$

Thus

$$\begin{aligned} x^{q_s} &= -(x^{2am} - 1)(x^{2am(2a-1)} + x^{2am(2a-3)} + \dots + x^{2ams}) \\ &= x^{am(2a+1)+2ams}(x^{2am} - 1)(x^{2am(2a-s)} + \dots + 1). \end{aligned}$$

Hence

$$x^{q_s-am(2a+1)-2ams} = x^{2am(2a\lambda+1-s)} - 1,$$

that is

$$x^{q_s+am(2a+1)-2ams} = x^{q_c} \quad (\text{where } c = 2a\lambda + 1 - s).$$

Thus the differences are

$$x^{q_s+ai+2arm} \text{ and } x^{q_s+am(2a\lambda+1-2s)+ai+2arm},$$

As in the preceding theorem one can show that

- (1) the  $x^{ai+2arm+q_s}$  are distinct for  $s$  fixed,  $i$  and  $r$  varying ;
- (2) the  $x^{q_s+am(2a\lambda+1-2s)+2arm}$  are distinct for  $s$  fixed,  $i$  and  $r$  varying.

Also, the sets (1) and (2) are disjoint ; for if

$$\begin{aligned} ai'+2ar'm &\equiv ai+2arm+am(2a\lambda+1-2s) & (\text{mod } 2am(2a\lambda+1)), \\ \text{then } (i-i') &\equiv m(2r-2r'+2a\lambda+1-2s) & (\text{mod } 2m(2a\lambda+1)), \\ i-i' &\equiv 0 & (\text{mod } m) \\ i &= i' \text{ (as before).} \end{aligned}$$

$$\text{Hence } 2(r-r')+(2a\lambda+1-2s) \equiv 0 \pmod{2(2a\lambda+1)}$$

which is impossible.

Let now  $q_u$  and  $q_w$  be any two elements from different residue classes (mod  $a$ ), that is

$$q_u - q_w \not\equiv 0 \pmod{a}.$$

Then as  $q_u$  and  $q_w$  range over the residue classes (mod  $a$ ) the corresponding differences

$$\begin{aligned} x^{ai+2arm+q_u}, & \quad x^{ai+2arm+q_u+am(2a\lambda+1-2u)} \\ x^{ai+2arm+q_w}, & \quad x^{ai+2arm+q_w+am(2a\lambda+1-2w)}, \end{aligned}$$

are all distinct.

For if not, there exist  $i, i', r, r'$  satisfying at least one of the following congruences :

- (1)  $ai+2arm+q_u \equiv ai'+2ar'm+q_w \pmod{2am(2a\lambda+1)}$
- (2)  $ai+2arm+q_u \equiv ai'+2ar'm+q_w+am(2a\lambda+1-2w)$  , ,
- (3)  $ai+2arm+q_u+am(2a\lambda+1-2u) \equiv ar'+2ar'm+q_w$  , ,
- (4)  $ai+2arm+q_u+am(2a\lambda+1-2u) \equiv ai'+2ar'm+q_w+am(2a\lambda+1-2w)$  , ,

All of these relations are impossible since

$$q_u - q_w \not\equiv 0 \pmod{a}.$$

Hence all the differences are distinct, and are  $2am(2a\lambda+1)$  in number for  $q_u - q_w$  ranging once over the residue classes (mod  $a$ ); therefore they range once over  $GF(v)$ , and  $\lambda$  times in all as the  $q_u - q_w$  cover the residue classes (mod  $a$ )  $\lambda$  times.

Series  $B$  and  $D$  (Sprott, 1954) are special cases of this theorem for  $a = 1$  and  $a = 2$  respectively.

*Example :* Let  $a = 4$ ,  $m = 1$ , and  $\lambda = 1$ ; thus  $v = 73$ . There is one set of 4 elements, namely

$$5^8 - 1 \equiv 5^0, \quad 5^{24} - 1 \equiv 5^{33}, \quad 5^{16} - 1 \equiv 5^6, \quad 5^{32} - 1 \equiv 5^7, \quad \text{all modulo } 73.$$

# SOME SERIES OF BALANCED INCOMPLETE BLOCK DESIGNS

The residue classes (mod 4) represented by the  $q$  are 0, 1, 2, and 3 (each more). Hence the design  $v = b = 73$ ,  $r = k = 9$ ,  $\lambda = 1$  can be constructed from the initial block

$$(5^0, 5^8, 5^{16}, 5^{24}, 5^{32}, 5^{40}, 5^{48}, 5^{56}, 5^{64}) \text{ or}$$

$$(1, 2, 4, 8, 16, 32, 64, 55, 37).$$

The last theorem is a special case of Theorem 4.1 in Sprott (1955).

## 4. SERIES 3

**Theorem 4.1:** If  $v = 2am(2a\lambda - 1) + 1 = p^n$  where  $p$  is a prime and if among the exponents  $q_s$ ,

$$x^{2ams} - 1 = x^{q_s} (s = 1, 2, \dots, a\lambda - 1),$$

where

the residue class of  $(j-1) \pmod{a}$  is represented  $\lambda$  times for all  $j \neq 1$ , while the residue class of 0 (mod  $a$ ) is represented  $\lambda-1$  times, then the design with parameters

$$v = 2am(2a\lambda - 1) + 1, \quad b = mv, \quad r = 2am\lambda, \quad k = 2a\lambda, \quad \lambda,$$

can be constructed from the initial blocks

$$(0, x^{ai}, x^{ai+2am}, \dots, x^{ai+2am(2a\lambda-2)})$$

where  $i$  ranges from 0 to  $m-1$  and  $x$  is a primitive element of  $GF(v)$ .

*Proof:* The differences not involving the zero element cover all non-zero elements of  $GF(v)$   $\lambda-1$  times from the preceding theorem, and all elements of the form  $x^u (u \neq 0 \pmod{a})$   $\lambda$  times. The elements in  $GF(v)$  of the form  $x^{aw}$  occur  $\lambda-1$  times among the preceding differences, and once among the differences involving the zero element. Thus the differences occur  $\lambda$  times in all and the design can be constructed by the first module theorem.

**Corollary 4.1:** Let  $a = 2$ . The resulting design is

$$v = 4m(4\lambda - 1) + 1, \quad b = mv, \quad r = 4m\lambda, \quad k = 4\lambda, \quad \lambda.$$

Here

$$x^{q_s} = x^{4ms} - 1 \quad (s = 1, 2, \dots, 2\lambda - 1);$$

among the  $q_s$  there must be  $\lambda$  odd powers and  $\lambda-1$  even powers of  $x$ .

**Corollary 4.2:** In Corollary 4.1 let  $\lambda = 1$ . The design is

$$v = 12m + 1, \quad b = mv, \quad r = 4m, \quad k = 4, \quad \lambda = 1.$$

The condition on the  $q_s$  simplifies to the requirement that  $x^{4m} - 1$  be an odd power of  $x$ . This is series  $F_1$  of Bose (1939).



## 5. SERIES 4

Theorem 5.1: If  $4u-1 = p^n$ , where  $p$  is a prime, then the design with parameters

$$v = 4u, \quad b = 2(4u-1), \quad r = 4u-1, \quad k = 2u, \quad \lambda = 2u-1,$$

can be constructed from the initial blocks

$$(0, x^0, x^2, \dots, x^{4(u-1)}) \quad (\infty, x, x^3, \dots, x^{4u-3}),$$

where  $x$  is a primitive element of  $GF(v-1)$ . This series of designs has two additional properties: (1) there are  $u-1$  blocks containing any three given elements: (2) the designs are all affine resolvable. (Bose (1942) discusses resolvability and affine resolvability.)

Proof: The differences arising from the first initial block are just those arising from the symmetric series contained in series  $C$  of Sprott (1954) with  $\lambda$  replaced here by  $u$ . Thus, the first initial block gives rise to a design

$$v = b = 4u-1, \quad r = k = 2u, \quad \lambda' = u.$$

The second initial block (without  $\infty$ ) contains all elements not contained in the first initial block, and is  $x$  times the initial block of the symmetric series of series  $B$  of Sprott (1954), with the  $\lambda$  of that series replaced here by  $u-1$ .

Thus the second initial block gives rise to the design

$$v = b = 4u-1, \quad r = k = 2u-1, \quad \lambda'' = u-1.$$

Using the notation of the second module theorem,

$$ns - \lambda = 4u-1 - (2u-1) = 2u = k = kt,$$

and

$$\lambda = (k-1)s = (k-1)t = 2u-1.$$

Hence condition (2) of the second module theorem is satisfied with  $s = t = 1$ , the differences occurring  $u$  times in one set of blocks and  $u-1$  times in the other, and hence  $\lambda = 2u-1$  times in all. Thus the design can be formed by the second module theorem. The design is obviously resolvable since the two initial blocks contain all elements and hence constitute one replication; also,  $b = v+r-1$ , and so the design is affine resolvable, any two blocks from different replications having  $k^2/v = u$  elements in common.

In the sub-design formed from the first initial block, any pair of elements occurs  $u$  times, and any element occurs  $2u$  times. Suppose any triplet  $a, b, c$  occurs  $c_{abc}$  times in this sub-design. Then exactly two of  $a, b, c$  occur in  $u - c_{abc}$  blocks, and exactly one of  $a, b, c$  occurs in  $2u - 2(u - c_{abc}) - c_{abc} = c_{abc}$  blocks. Hence the total number of blocks in this sub-design containing  $a, b$ , or  $c$ , is  $c_{abc} + 3(u - c_{abc}) + 3c_{abc} = c_{abc} + 3u$ , and so the number of blocks of the sub-design not containing  $a, b$ , or  $c$ , is  $4u-1 - (c_{abc} + 3u) = u-1 - c_{abc}$ . Such blocks give rise to blocks in the second sub-design (arising from the second initial block) which contain all of  $a, b$ , and  $c$ ; thus the number of blocks in the complete design containing  $a, b$ , and  $c$  is  $c_{abc} + (u-1 - c_{abc}) = u-1$  for any triplet  $a, b, c$ . Also,  $\infty$  occurs in all blocks formed from the second initial block, and so occurs  $u-1$  times with all pairs of elements.

# SOME SERIES OF BALANCED INCOMPLETE BLOCK DESIGNS

## 6. SERIES 5

Theorem 6.1: If  $2k-1 = p^n$ , where  $p$  is a prime, then the design with parameters

$$v = 2k, \quad b = 4(2k-1), \quad r = 2(2k-1), \quad k, \quad \lambda = 2(k-1).$$

can be constructed from the initial blocks

$$(0, x^i, x^{i+2}, \dots, x^{i+2k-4}), \quad (\infty, x^{i+1}, x^{i+3}, \dots, x^{i+2k-3}),$$

where  $x$  is a primitive element of  $GF(v-1)$  and  $i = 0, 1$ . Further, every triplet occurs  $k-2$  times and the design is resolvable.

*Proof:* Apply the second module theorem with  $s = t = 2$ . Because of  $ns - \lambda = 2(v-1)2 - 2(k-1) = 4k-2-2k+2 = kt$ , and  $\lambda = 2(k-1) = s(k-1)$ , condition (2) of the second module theorem is satisfied. The second initial block is  $x$  times that a series  $A$  (Sprott 1954) with  $m = 2$ , and  $k$  replaced here by  $k-1$ ; thus with  $\infty$  deleted the differences occur  $k-2$  times each and the resulting sub-design is

$$v = 2k-1, \quad b = 2v, \quad r = 2k, \quad k-1, \quad \lambda' = k-2.$$

The first set of initial blocks contain all elements not contained in the second set, and are the initial blocks of series 1 of this paper with  $m = 2$  and  $\lambda$  replaced by  $k$ . Thus they give rise to a second sub-design with parameters

$$v = 2k-1, \quad b = 2v, \quad r = 2k, \quad k, \quad \lambda'' = k.$$

Thus the differences occur  $\lambda = 2k-2$  times in all, and the design can be constructed by the second module theorem.

As in Theorem 5.1, it can be proved that any three elements are continued in  $(k-2)$  blocks and that the design is resolvable.

Possible practical applications of designs having the property that any three elements occur together in  $\delta$  blocks were considered by Calvin (1954); he constructed several designs which are special cases of the preceding two theorems.

*Example 6.1:* In series (4) let  $u = 2$ ; the resulting design has parameters  $v = 8, \quad b = 14, \quad r = 7, \quad k = 4, \quad \lambda = 3$ .

The initial blocks are  $(0, 3^0, 3^2, 3^4), (\infty, 3, 3^3, 3^5)$ .

$(0, 1, 2, 4)$	$(\infty, 3, 6, 5)$	$(4, 5, 6, 1)$	$(\infty, 0, 3, 2)$
$(1, 2, 3, 5)$	$(\infty, 4, 0, 6)$	$(5, 6, 0, 2)$	$(\infty, 1, 4, 3)$
$(2, 3, 4, 6)$	$(\infty, 5, 1, 0)$	$(6, 0, 1, 3)$	$(\infty, 2, 5, 4)$
$(3, 4, 5, 0)$	$(\infty, 6, 2, 1)$		

The blocks are written in pairs, each pair constituting a complete replication. Any two blocks from different replications have exactly two elements in common. Every set of three elements occurs exactly once.

*Example 6.2 :* In series (5) let  $k = 7$ ; the resulting design has parameters  $v = 14$ ,  $b = 52$ ,  $r = 26$ ,  $k = 7$ ,  $\lambda = 12$ .

The initial blocks are

$$\begin{array}{ll} (0,1,4,3,12,9,10) & (\infty,2,8,6,11,5,7) \\ (0,2,8,6,11,5,7) & (\infty,1,4,3,12,9,10), \end{array}$$

each pair constituting a complete replication.

Thus the design is resolvable and can be formed by addition (mod 13). Every triplet occurs five times.

#### REFERENCES

- BOSE, R. C. (1939) : On the construction of balanced incomplete block designs. *Ann. Eugen.*, **9**, 353-399.  
 ——— (1942) : On some new series of balanced incomplete block designs. *Bull. Cal. Math. Soc.*, **34**, 17-31.  
 ——— (1942) : A note on the resolvability of balanced incomplete block designs. *Sankhyā*, **6**, 105-110.  
 CALVIN, L. D. (1954) : Doubly balanced incomplete block designs for experiments in which effects are correlated. *Biometrics*, **10**, 61-88.  
 SPROTT, D. A. (1954) : A note on balanced incomplete block designs. *Canadian J. Math.*, **6**, 341-346.  
 ——— (1955) : Some series of partially balanced incomplete block designs. (to be published in the *Canadian J. Math.*)

*Paper received : April, 1955*

# MISCELLANEOUS

## THE CONCEPT OF ASYMPTOTIC EFFICIENCY

By D. BASU

*Indian Statistical Institute, Calcutta*

### 1. SUMMARY

Partly of an expository nature this note brings out the fact that an estimator, though asymptotically much less efficient (in the classical sense) than another, may yet have much greater probability concentration (as defined in this article) than the latter.

### 2. DEFINITIONS

Let  $\{X_i\}$ ,  $i = 1, 2, \dots$  be an infinite sequence of independent and identically distributed random variables whose common distribution function  $F$  is known to belong to a family  $\Omega$  of one dimensional distribution functions. Let  $\mu = \mu(F)$  be a real valued functional defined on  $\Omega$ . By an estimator  $T = \{t_n\}$  of  $\mu$  we mean a sequence of real valued measurable functions of  $\{X_i\}$ , where  $t_n$  is a function of  $X_1, X_2, \dots, X_n$  only ( $n = 1, 2, \dots$ ). The estimator  $T$  is said to be an asymptotically normal estimator of  $\mu$  if there exists a sequence  $\{\sigma_n(F)\}$  of positive numbers such that as  $n \rightarrow \infty$

$$\{t_n - \mu(F) / \sigma_n(F)\} \implies N(0, 1) \quad \text{for all } F \in \Omega$$

where  $\implies$  stands for convergence in law and  $N(0, 1)$  for the standard normal variable. The sequence  $\{\sigma_n(F)\}$  is called the asymptotic standard deviation of  $T$ . A necessary and sufficient condition in order that  $\{\sigma_n(F)\}$  and  $\{\sigma'_n(F)\}$  may both be called the asymptotic standard deviation of  $T$  is

$$\lim_{n \rightarrow \infty} \{\sigma_n(F) / \sigma'_n(F)\} \equiv 1 \quad \text{for all } F \in \Omega.$$

A necessary and sufficient condition in order that the asymptotically normal estimator  $T$  is also consistent is

$$\lim_{n \rightarrow \infty} \sigma_n(F) \equiv 0 \quad \text{for all } F \in \Omega.$$

Let  $\mathcal{I}$  be the family of all consistent asymptotically normal estimators of  $\mu$ . We consider only the space  $\mathcal{I}$ .



## 3. THE PARTIAL ORDER OF EFFICIENCY

Two elements  $T$  and  $T'$  of  $\mathcal{J}$  are said to be equally efficient (or equivalent) if they have the same asymptotic s.d.s, i.e. if

$$\lim_{n \rightarrow \infty} \{\sigma_n(F)/\sigma'_n(F)\} \equiv 1 \quad \text{for all } F \in \Omega \quad \dots (3.1)$$

where  $\{\sigma_n(F)\}$  and  $\{\sigma'_n(F)\}$  are the corresponding asymptotic s. d.'s.

It is easily verified that the above equivalence relation is reflexive, symmetric, and transitive.

If 
$$\limsup_{n \rightarrow \infty} \{\sigma_n(F)/\sigma'_n(F)\} \leq 1 \quad \text{for all } F \in \Omega$$

and 
$$\liminf_{n \rightarrow \infty} \{\sigma_n(F)/\sigma'_n(F)\} < 1 \quad \text{for some } F \in \Omega$$

then we say that  $T$  is more efficient than  $T'$  and write  $T' \supset T$ . It is easily seen that the relation  $\supset$  induces a partial order on  $\mathcal{J}$ .

It is known that there do not exist a maximal element in  $\mathcal{J}$  with respect to the partial order  $\supset$ , i.e. there do not exist any element  $T \in \mathcal{J}$  which is either equivalent to or more efficient than any alternative  $T' \in \mathcal{J}$ . As a matter of fact it has been demonstrated (LeCam, 1953) how given any  $T \in \mathcal{J}$  we can always find a  $T' \in \mathcal{J}$  such that  $T' \supset T$ .

## 4. THE PARTIAL ORDER OF CONCENTRATION

The estimator  $T = \{t_n\}$  of  $\mu$  is consistent if for all  $\epsilon > 0$  and  $F \in \Omega$

$$p_n(\epsilon, F) = P\{|t_n - \mu| > \epsilon | F\} \rightarrow 0 \text{ as } n \rightarrow \infty.$$

If we work with the simple loss function that is zero or one according as the error in the estimate is  $\leq \epsilon$  or  $> \epsilon$  then  $p_n(\epsilon, F)$  is the risk (or expected loss) when the estimator is used with observations on  $X_1, X_2, \dots, X_n$  only.

The rapidity with which  $p_n(\epsilon, F) \rightarrow 0$  may be considered to be a measure of the asymptotic accuracy or concentration of  $T$ . This motivates the following definition of a partial order on  $\mathcal{J}$  (and as a matter of fact on the wider family of all consistent estimators of  $\mu$ ).

*Definition:* The estimator  $T$  with the associated sequences of risk functions  $p_n(\epsilon, F)$  is said to have greater concentration than  $T'$  with the associated sequences  $p'_n(\epsilon, F)$  if, for all  $\epsilon > 0$  and  $F \in \Omega$ ,

$$\limsup_{n \rightarrow \infty} \{p_n(\epsilon, F)/p'_n(\epsilon, F)\} \leq 1$$

with the limit inferior being  $< 1$  for some  $\epsilon > 0$  and some  $F \in \Omega$ . We then write  $T > T'$ .

# THE CONCEPT OF ASYMPTOTIC EFFICIENCY

Intuitively it may seem reasonable to expect that  $T \supset T'$  implies  $T > T'$ . That this is not so is demonstrated in the next section. An example is given where

$$\lim_{n \rightarrow \infty} \frac{\sigma_n(F)}{\sigma'_n(F)} \equiv 0 \quad \text{for all } F \in \Omega, \quad \dots (4.1)$$

$$\text{whereas} \quad \lim_{n \rightarrow \infty} \frac{p_n(\epsilon, F)}{p'_n(\epsilon, F)} \equiv \infty \quad \text{for all } \epsilon > 0 \text{ and } F \in \Omega. \quad \dots (4.2)$$

## 5. AN EXAMPLE

Let each of the  $X_i$ 's be  $N(\mu, 1)$ , the problem being to estimate  $\mu$ .

$$\text{Let} \quad \bar{X}_n = \sum_1^n X_i/n \text{ and } S_n = \sum_1^n (X_i - \bar{X}_n)^2.$$

Then  $\bar{X}_n$  and  $S_n$  are mutually independent random variables and the distribution of  $S_n$  is independent of  $\mu$ . Let  $a_n$  be the upper  $100/n$  % point of  $S_n$  and let

$$H_n = \begin{cases} 0 & \text{if } S_n \leq a_n, \\ 1 & \text{if } S_n > a_n. \end{cases}$$

Now let

$$T = \{t_n\}$$

where

$$t_n = (1 - H_n)\bar{X}_n + n H_n$$

and

$$T' = \{t'_n\}$$

where

$$t'_n = \bar{X}_{[ \sqrt{n} ]}$$

(By  $[x]$  we mean the largest integer not exceeding  $x$ .)

Since

$$P(H_n = 0) = 1 - \frac{1}{n} \rightarrow 1,$$

it follows (vide Cramér, p. 254) that  $\sqrt{n}(t_n - \mu) = \sqrt{n}(\bar{X}_n - \mu) + \sqrt{n} H_n(n - \bar{X}_n)$

$$\implies N(0, 1)$$

when  $\mu$  is the true mean.

Hence,  $T \in \mathcal{Z}$  with asymptotic s.d. =  $\{n^{-1/2}\}$ . Also  $T' \in \mathcal{Z}$  with asymptotic s.d. =  $\{n^{-1/4}\}$ .

Therefore (4.1) is satisfied. Again, since  $\bar{X}_n$  is independent of  $H_n$  it follows that, for every  $n > \mu + \varepsilon$ ,

$$\begin{aligned} P(|t_n - \mu| > \varepsilon | \mu) &= P(H_n = 0) P(|\bar{X}_n - \mu| > \varepsilon | \mu) + P(H_n = 1) \\ &= \frac{1}{n} + o\left(\frac{1}{n}\right) \end{aligned}$$

because  $P(|\bar{X}_n - \mu| > \varepsilon | \mu) = o\left(\frac{1}{n}\right)$ , as may be easily verified.

Whereas  $P(|t'_n - \mu| > \varepsilon | \mu) = o\left(\frac{1}{n}\right)$

Therefore (4.2) also is satisfied.

It may be noted that in the example given the s.d. of  $t_n$  is not asymptotically equal to the asymptotic s.d. of  $T$ . But this can be easily arranged to be true by, say, taking  $a_n$  for the upper  $100/n^4$  % point of  $S_n$ .

#### REFERENCES

- CRAMÉR, H. (1946) : *Mathematical Methods of Statistics*, Princeton University Press.  
 LECAM, L. (1953) : *On Some Asymptotic Properties of Maximum Likelihood Estimates and Related Bayes Estimates*, University of California Press.

*Paper received : January, 1955.*

# A NOTE ON THE DETERMINATION OF OPTIMUM PROBABILITIES IN SAMPLING WITHOUT REPLACEMENT

By DES RAJ  
Indian Statistical Institute, Calcutta

## 1 INTRODUCTION

It is now well-known that the use of varying probabilities in selecting a sample may bring about considerable reduction in the sampling variance of the estimate as compared to the case when the units are selected with equal probabilities. This technique was first suggested by Hansen and Hurwitz (1943) who considered a design in which one first stage unit is selected with probability proportional to size within each stratum. Horvitz and Thompson (1952) generalised it to the selection of  $n$  units without replacement within strata. Their estimator of the population (or stratum) total is

$$\hat{y} = \sum_1^n \frac{y_i}{\pi_i} \quad \dots (1.1)$$

with variance

$$V(\hat{y}) = \sum_1^N \frac{y_i^2}{\pi_i} + 2 \sum' \pi_{ij} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} - Y^2 \quad \dots (1.2)$$

where  $\pi_{ij}$  = probability with which two units  $u_i$  and  $u_j$  enter the sample,  
 $\pi_i$  = probability that the unit  $u_i$  is selected in the sample,  
 $y_i$  = value of the unit  $u_i$  for the character  $y$ ,

and  $\Sigma'$  denotes summation over the  ${}^N C_2$  pairs of units.

An important problem then arises: How should the sample be selected so that the variance of the estimate is made smallest? It is easy to see that the estimator has zero variance if  $\pi_i \propto y_i$ . This result is not of practical interest because if the  $y_i$  were known in advance, the sample would be unnecessary. The result, however, suggests that if the values of the units for a known auxiliary character  $x$  are reasonably proportional to  $y$ , it would be advantageous to select the sample so that  $\pi_i \propto x_i$ .

Now there may be several methods of drawing such a sample. Each such method will lead to certain  $\pi_{ij}$  whose values are going to affect the variance of the estimate. It is then of interest to choose a sampling scheme from all such schemes such that the  $\pi_{ij}$  associated with this scheme should minimise  $V(\hat{y})$  given by (1.2). The object of this note is to obtain such an optimum scheme.

## 2. SOLUTION OF THE PROBLEM

Considering the practically useful case of  $n = 2$  (and the method is unlikely to be convenient for larger sample sizes), the problem consists in the determination of  ${}^N C_2$  probabilities of selection of pairs,  $\pi_{ij}$ , such that

$$\pi_{ij} \geq 0, \quad \dots (2.1)$$

$$\sum_{j \neq i} \pi_{ij} = \pi_i \quad (i = 1, 2, \dots, N) \quad \dots (2.2)$$

and

$$\sum' \pi_{ij} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} \quad \dots (2.3)$$

is minimised.



Stated thus, this is a familiar problem in linear programming. But the difficulty involved is that the coefficients of  $\pi_{ij}$  in (2.3) are unknown. We shall make the assumption that

$$y = \alpha + \beta x \quad \dots (2.4)$$

i.e. the relation between  $y$  and  $x$  is a straight line (and the method is unlikely to be useful if the relation is not linear). We shall, however, not assume any knowledge of the actual values of  $\alpha$  and  $\beta$ . The quantity to be minimised then is found to be

$$\sum' \frac{\pi_{ij}}{\pi_i \pi_j} \quad \dots (2.5)$$

This result follows from the observation that

$$\begin{aligned} \sum' \pi_{ij} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} &= \frac{1}{2} \sum_{i=1}^N \sum_{j \neq i}^N \frac{\pi_{ij}}{\pi_i \pi_j} (\alpha^2 + \alpha \beta x_i + \alpha \beta x_j + \beta^2 x_i x_j) \\ &= \frac{\alpha^2}{2} \sum \sum \frac{\pi_{ij}}{\pi_i \pi_j} + \frac{\left( \sum_1^N x_i \right)}{4} \alpha \beta \left( \sum \sum \frac{\pi_{ij}}{\pi_i} + \sum \sum \frac{\pi_{ij}}{\pi_j} \right) \\ &\quad + \frac{\left( \sum_1^N x_i \right)^2}{8} \beta^2 \sum \sum \pi_{ij} \\ &= \frac{\alpha^2}{2} \sum \sum \frac{\pi_{ij}}{\pi_i \pi_j} + \frac{N}{2} \alpha \beta \left( \sum_1^N x_i \right) + \frac{\left( \sum_1^N x_i \right)^2}{4} \beta^2 \end{aligned}$$

since

$$\pi_i = \frac{2x_i}{\sum_1^N x_i}, \pi_j = \frac{2x_j}{\sum_1^N x_j}, \sum_1^N \pi_i = \sum_1^N \pi_j = 2,$$

$$\sum_{j \neq i} \pi_{ij} = \pi_i, \sum_{i \neq j} \pi_{ij} = \pi_j.$$

Then the problem reduces to the determination of  $\pi_{ij}$  such that

$$\left. \begin{aligned} \pi_{ij} &\geq 0, \\ \sum_{j \neq i} \pi_{ij} &= \pi_i \quad (i = 1, 2, \dots, N) \\ \sum' \pi_{ij} / (\pi_i \pi_j) &\text{ is minimised.} \end{aligned} \right\} \quad \dots (2.6)$$

and

# OPTIMUM PROBABILITIES IN SAMPLING WITHOUT REPLACEMENT

As stated before, this is a problem in linear programming and can be solved by the simplex method given in Charnes and others (1953).

## 3. SELECTION OF THE SAMPLE

With regard to the actual procedure of drawing the sample ensuring  $\pi_i \propto x_i$ , various methods are now available. Horvitz and Thompson (1952) have suggested two methods which are of limited applicability. Narain's (1951) method requires the solution of an equation by graphical numerical methods. Yates and Grundy (1953) obtain revised size-measures (by an iterative process) and obtain the sample by selecting the first unit with probabilities proportional to the revised sizes and the second unit with probabilities proportional to the remaining sizes. Goodman and Kish (1950) have devised a very convenient method of selection by cumulating the sizes and drawing a systematic sample from the cumulated sizes, the  $\pi_{ij}$  depending on the order in which the units are listed. One characteristic common to all these methods is that they seek to arrive at some  $\pi_{ij}$  which may by no means be optimum. The method of drawing the sample, considered in this note, is however very simple. Out of the totality of  ${}^N C_2$  groups (of two units each), one has to select one group with given probabilities assigned in an optimum way.

## 4. AN ILLUSTRATION

As an illustration of the practical utility of the method we consider the three populations  $A$ ,  $B$  and  $C$  given by Yates and Grundy (1953). These populations were deliberately chosen by them as being more extreme than will normally be encountered in practice. The object is to estimate the population total by selecting two units with probabilities of inclusion  $\pi_i$  proportional to the following  $p_i$ .

unit	$p$
1	0.1
2	0.2
3	0.3
4	0.4

We have to find  $\pi_{ij}$  such that

$$\begin{aligned} \pi_{ij} &\geq 0, \\ \pi_{12} + \pi_{13} + \pi_{14} &= 0.2, \quad \pi_{21} + \pi_{23} + \pi_{24} = 0.4, \\ \pi_{13} + \pi_{32} + \pi_{34} &= 0.6, \quad \pi_{41} + \pi_{42} + \pi_{43} = 0.8, \end{aligned}$$

$$\text{and } G = 12.5\pi_{12} + 8.3333\pi_{13} + 6.25\pi_{14} + 4.1667\pi_{23} + 3.125\pi_{24} + 2.0833\pi_{34}$$

is minimised.

The optimum assignment of  $\pi_{ij}$ , obtained by the simplex method, is given in Table 1 below.

TABLE 1. OPTIMUM ASSIGNMENT OF  $\pi_{ij}$

$u_i \backslash u_j$	1	2	3	4	total
1	—	0.0	0.0	0.2	0.2
2	0.0	—	0.2	0.2	0.4
3	0.0	0.2	—	0.4	0.6
4	0.2	0.2	0.4	—	0.8
total	0.2	0.4	0.6	0.8	2.0

Yates and Grundy's assignment of probabilities is given in Table 2. In this case revised size-measures, based on three successive approximations, were used.

TABLE 2. YATES AND GRUNDY'S ASSIGNMENT OF  $\pi_{ij}$ 

$u_i$	$u_j$	1	2	3	4	total
1	1	—	.032	.059	.113	.204
2	1	.032	—	.122	.246	.400
3	1	.059	.122	—	.428	.609
4	1	.113	.246	.428	—	.787
total	1	.204	.400	.609	.787	2.000

Denoting by  $V_{opt}$  and  $V_{YG}$  the variances of the estimate  $\hat{y}$ , when the  $\pi_{ij}$  are taken from Tables 1 and 2 respectively, the following results (given in Table 3) are obtained. It may be observed that the relation between  $y$  and  $x$  may be assumed to be approximately linear for populations  $A$  and  $B$  but not so for population  $C$ . It is also found that the set of  $\pi_{ij}$  minimising (2.5) is the same as the set minimising (2.3) for populations  $A$  and  $B$  while it is not so for population  $C$ .

TABLE 3. COMPARISON OF SAMPLING VARIANCES

	$V_{YG}$	$V_{opt}$	% reduction in variance
population $A$	.323	.200	38.1
population $B$	.269	.200	25.7
population $C$	.057	.100	-75.4

### 5. CONCLUSION

These results show that if the relation between  $y$  and  $x$  can be assumed to be linear, the optimum assignment of  $\pi_{ij}$ , suggested in this note, can bring about marked reduction in variance as compared to the usual method of assignment. But this refinement is conveniently obtainable only when the number of units in a stratum is small. The method is expected to be of considerable use if the population is stratified to the maximum extent possible and the selection of two first stage units within strata is considered to be adequate.

### REFERENCES

- CHARNES, A., COOPER, W. W. AND HENDERSON, A. (1953): "An Introduction to Linear Programming", John Wiley & Sons, New York.
- GOODMAN, R. AND KISH, L. (1950): Controlled selection—a technique in probability sampling. *J. Amer. Stat. Ass.*, **45**, 350-372.
- HANSEN, M. H. AND HURWITZ, W. N. (1943): On the theory of sampling from finite populations. *Ann. Math. Stat.*, **14**, 333-362.
- HORVITZ, D. G. AND THOMPSON, D. J. (1952): A generalisation of sampling without replacement from a finite universe. *J. Amer. Stat. Ass.*, **47**, 663-685.
- NARAIN, R. D. (1951): On sampling without replacement with varying probabilities. *J. Ind. Soc. Agri. Stat.*, **3**, 169-174.
- YATES, F. AND GRUNDY, P. M. (1953): Selection without replacement from within strata with probabilities proportionate to size. *J. Roy. Stat. Soc., B*, **15**, 253-261.

Paper received : May, 1955.



# SANKHYĀ

## THE INDIAN JOURNAL OF STATISTICS

Edited by : P. C. MAHALANOBIS

VOL. 17, PART 3

DECEMBER

1956

### ALMOST UNBIASSED ESTIMATES OF FUNCTIONS OF FREQUENCIES

By J. B. S. HALDANE

*University College, London*

If a sample of  $n$  contains  $a$  individuals with a character  $A$ . then, provided sampling is random, and provided  $a$  is not zero or equal to  $n$ ,  $\frac{a}{n}$  is a satisfactory estimate of the probability  $p$  that the next individual sampled will have the character  $A$ , whether or not we regard the process of sampling as the repetition of an experiment, or the selection of individuals from a pre-existing population. It is the maximum likelihood estimate, and also an unbiased estimate.

But if we are concerned with a function  $F(p)$  of this probability, then  $F\left(\frac{a}{n}\right)$  is the maximum likelihood estimate of  $F(p)$ . But it is never an unbiased estimate. Let us take an example to show that this is of practical importance.

De Winton and Haldane (1933) pollinated "oak-leaved" recessive *Primula sinensis* with pollen from heterozygotes. They obtained 154 normal plants and 145 with "oak" leaves. Since characters which can be scored on seeds, for example, Mendel's character of yellow vs. green cotyledons, frequently do not deviate significantly from their expected numbers on samples of over 100,000, deviations from equality may be used to estimate the viability of the recessive compared with the normal type, since a number of seeds do not germinate, and seedlings die before producing leaves which can be scored. The maximum likelihood estimate of the viability is  $\frac{145}{154} = 0.9416 \pm 0.1090$ . The unbiased estimate, as we shall see, is  $\frac{145}{155} = 0.9355$ .

The bias is only 5.6% of the standard error, and negligible. But the same authors give data on 75 such segregations. Had they all been based on such small numbers the mean estimated viability would have been in excess by about half its standard error. Fortunately 36 of the samples considered contained over 1000 plants, but even so De Winton and Haldane's general conclusions require slight modification. In this case if  $p$  be the frequency of normal plants the relative viability of abnormal plants is  $p^{-1}-1$ . We thus require an unbiased estimate of  $p^{-1}$ .



Now we can obtain unbiased estimate of  $p^m$ , where  $m$  is a positive integer less than  $n$ . In this case it is well-known and easily proved that

$$\left[ \frac{a(a-1)(a-2)\dots(a-m+1)}{n(n-1)(n-2)\dots(n-m+1)} \right] = p^m. \quad \dots (1)$$

That is to say  $\frac{a(a-1)(a-2)\dots(a-m+1)}{n(n-1)(n-2)\dots(n-m+1)}$  or  $\frac{a!(n-m)!}{n!(a-m)!}$  is an unbiased estimate of  $p^m$ . It is also an efficient estimate, since as  $n$  tends to infinity it is asymptotically equivalent to the maximum likelihood estimate  $\left(\frac{a}{n}\right)^m$ . Can we find functions  $f(a, m)$  and  $f(n, m)$  such that  $\left[\frac{f(a, m)}{f(n, m)}\right]^m$  is an approximately unbiased estimate of  $p^m$  when  $m$  is not a positive integer? If  $m$  and  $a$  are positive integers

$$\begin{aligned} \ln[f(a, m)] &= m^{-1} \left[ m \ln a + \sum_{r=1}^{m-1} \ln(1-ra^{-1}) \right] \\ &= \ln a - m^{-1} \sum_{i=1}^{\infty} \left[ i^{-1} a^{-i} \sum_{r=1}^{m-1} r^i \right] \\ &= \ln a - m^{-1} \left[ \frac{m(m-1)}{2a} + \frac{m(m-1)(2m-1)}{12a^2} + \frac{m^2(m-1)^2}{12a^3} + \right. \\ &\quad \left. + \frac{m(m-1)(2m-1)(3m^2-3m-1)}{120a^4} + \frac{m^2(m-1)^2(2m^2-2m-1)}{60a^5} + \dots \right] \\ &= \ln a - \frac{m-1}{2a} \left[ 1 + \frac{2m-1}{6a} + \frac{m(m-1)}{6a^2} + \frac{(2m-1)(3m^2-3m-1)}{60a^3} + \right. \\ &\quad \left. + \frac{m(m-1)(2m^2-2m-1)}{30a^4} + \dots \right]. \end{aligned} \quad \dots (2)$$

It can easily be shown that this is also the expansion of  $\left[\frac{a!}{(a-m)!}\right]^{m-1}$  when  $m$  is a negative integer. Hence

$$\begin{aligned} f(a, m) &= a - \frac{1}{2}(m-1) \left[ 1 + \frac{m+1}{12a} \left( 1 + \frac{m-1}{2a} + \frac{73m^2-120m+23}{240a^2} + \right. \right. \\ &\quad \left. \left. + \frac{(m-1)(33m^2-40m-17)}{160a^3} + \dots \right) \right] + O(a^{-5}). \quad \dots (3) \end{aligned}$$

# ALMOST UNBIASED ESTIMATES OF FUNCTIONS OF FREQUENCIES

Since  $f(a, 1) = a$ ,  $m-1$  must be a factor of all terms after the first, and since  $f(a, -1) = a+1$ ,  $m+1$  must be a factor of all terms after the second. If  $a=0$ ,  $f(a, m) = 0$  when  $m$  is positive, and is undefined when  $m$  is negative. The following special cases are useful:

$$f(a, \frac{1}{2}) = a + \frac{1}{4} + \frac{1}{32a} - \frac{1}{128a^2} - \frac{5}{2048a^3} + \frac{23}{4096a^4} + \dots \quad \dots (4)$$

$$f(a, 0) = a + \frac{1}{2} + \frac{1}{24a} - \frac{1}{48a^2} + \frac{23}{5760a^3} + \frac{17}{3840a^4} + \dots \quad \dots (5)$$

in each case provided  $a$  is a positive integer.  $f(a, m) = \left[ \frac{\Gamma(a+1)}{\Gamma(a-m+1)} \right]^{m-1}$

This function has a pole when  $a$  is any negative integer. Thus in the  $a^{-1}$  complex plane there is a sequence of poles along the real negative axis, of which zero is the limit point. Thus the series (3), like so many connected with the gamma function, has no circle of convergence, though it terminates when  $m$  is a positive integer. It is an asymptotic expansion.

$\ln f(a, 0) = \frac{d}{da} \ln \Gamma(a+1)$ , whose asymptotic expansion is the well known series.

$$\ln f(a, 0) = \ln a + \frac{1}{2a} - \frac{1}{12a^2} + \frac{1}{120a^4} = \ln a + \frac{1}{2a} + \sum_{r=1}^{\infty} \frac{B_r}{2ra^{2r}} \quad \dots (6)$$

where  $B_r$  is the  $r$ -th Bernoulli number.

I now define an almost unbiased estimate as an estimate whose bias, as the sample number  $n$  increases, tends to zero more rapidly than any negative power of  $n$ . In a former paper (Haldane, 1953) I used this terminology for an estimate whose bias tends to zero with  $n^{-2}$  or quicker. I think that the stronger definition is more appropriate.

I shall show that  $\left[ \frac{f(a, m)}{f(n, m)} \right]^m$ , where  $f(a, m)$  is the series whose first six terms are given by (3), is an almost unbiased estimate of  $p^m$  for all values of  $m$  less than  $n$  (though in fact the series is of little value when  $m$  is much larger than  $pn$ ). Similarly I conjecture that  $\ln f(a, 0) - \ln f(n, 0)$  is an almost unbiased estimate of  $\ln p$ . Its bias is certainly of order  $n^{-4}$  or less.

I first consider the case where  $m$  is a negative integer  $-k$ .

$$\text{Then} \quad [f(a, -k)]^{-k} = \frac{a!}{(a+k)!}$$

The probability of obtaining just  $a$  out of  $n$  with the character  $A$  is  $P_a = \binom{n}{a} p^a (1-p)^{n-a}$ . So

$$\begin{aligned} \mathcal{E} \left[ \frac{a!}{(a+k)!} \right] &= \sum_{a=0}^n \frac{a! P_a}{(a+k)!} \\ &= p^{-k} \frac{n!}{(n+k)!} \sum_{a=0}^n \binom{n+k}{a+k} p^{a+k} (1-p)^{n-a} \\ &= p^{-k} \frac{n!}{(n+k)!} \left[ 1 - \sum_{r=0}^{k-1} \binom{n+k}{r} p^r (1-p)^{n+k-r} \right] \\ &= p^{-k} \frac{n!}{(n+k)!} \left[ 1 - (1-p)^{n+1} \sum_{r=0}^{k-1} \binom{n+k}{r} p^r (1-p)^{k-r-1} \right]. \end{aligned}$$

$$\begin{aligned} \text{Now } \sum_{r=0}^{k-1} \binom{n+k}{r} p^r (1-p)^{k-r-1} &= \binom{n+k}{k-1} p^{k-1} + \binom{n+k}{k-2} p^{k-2} (1-p) + \dots \\ &\quad + \binom{n+k}{1} p (1-p)^{k-2} + (1-p)^{k-1}. \end{aligned}$$

That is to say it is a polynomial in  $n$  of degree  $k-1$ . Hence

$$\mathcal{E} \left[ \left\{ \frac{f(a, -k)}{f(n, -k)} \right\}^{-k} \right] = p^{-k} \left[ 1 - O\{(1-p)^n n^{k-1}\} \right]. \quad \dots (7)$$

That is to say the bias tends to zero more rapidly than any negative power of  $n$  as  $n$  becomes large. This result is equivalent to Bhattacharyya's (1954) result (1.9). I think that Bhattacharyya's methods could probably be applied to the results of this paper with an appreciable gain in rigour. Though I venture to prefer my definition of an almost unbiased estimate to his. The way in which the bias diminishes as  $n$  increases accords with the well-known property of the sum of an asymptotic expansion.

Now let  $a = pn + \theta$ . Then from the well-known expressions for the moments of a binomial distribution

$$\begin{aligned} \mathcal{E}(\theta) &= 0, \quad \mathcal{E}(\theta^2) = np(1-p), \quad \mathcal{E}(\theta^3) = np(1-p)(1-2p), \\ \mathcal{E}(\theta^4) &= 3n^2 p^2 (1-p)^2 + np(1-p)(1-6p+6p^2), \\ \mathcal{E}(\theta^5) &= 10n^2 p^2 (1-p)^2 (1-2p) + O(n^2), \quad \mathcal{E}(\theta^6) = 15n^3 p^3 (1-p)^3 + O(n^2), \\ \mathcal{E}(\theta^7) &= O(n^3) \text{ etc.} \end{aligned}$$

$$\text{Also } [f(a, m)]^m = a^m - \frac{1}{2} m(m-1) a^{m-1} + \frac{1}{24} m(m-1)(m-2)(3m-1) a^{m-2} \dots$$

So  $\mathcal{E}\{[f(a, m)]^m\}$

$$\begin{aligned} &= p^m n^m \mathcal{E} \left[ \left( 1 + \frac{\theta}{pn} \right)^m \right] - \frac{1}{2} m(m-1) p^{m-1} n^{m-1} \mathcal{E} \left[ \left( 1 + \frac{\theta}{pn} \right)^{m-1} \right] + \\ &\quad + \frac{1}{24} m(m-1)(m-2)(3m-1) p^{m-2} n^{m-2} \mathcal{E} \left[ \left( 1 + \frac{\theta}{pn} \right)^{m-2} \right] + \dots \end{aligned}$$

# ALMOST UNBIASED ESTIMATES OF FUNCTIONS OF FREQUENCIES

$$\begin{aligned}
 &= p^m n^m \left[ 1 + \frac{m(m-1)\mathcal{E}(\theta^2)}{2p^2 n^2} + \frac{m(m-1)(m-2)\mathcal{E}(\theta^3)}{6p^3 n^3} + \right. \\
 &\quad \left. + \frac{m(m-1)(m-2)(m-3)\mathcal{E}(\theta^4)}{24p^4 n^4} \right] - \frac{1}{2} m(m-1) p^{m-1} n^{m-1} \left[ 1 + \frac{(m-1)(m-2)\mathcal{E}(\theta^2)}{2p^2 n^2} \right] + \\
 &\quad + \frac{1}{24} m(m-1)(m-2)(3m-1) p^{m-2} n^{m-2} [1 + \dots] + O(n^{m-3}) \\
 &= p^m n^m + p^m n^{m-1} \left[ \frac{1}{2} m(m-1)(p^{-1} - 1) - \frac{1}{2} m(m-1)p^{-1} \right] + p^m n^{m-2} \left[ \frac{1}{6} m(m-1)(m-2) \times \right. \\
 &\quad \times (p^{-2} - 3p^{-1} + 2) + \frac{1}{8} m(m-1)(m-2)(m-3)(p^{-2} - 2p^{-1} + 1) - \\
 &\quad \left. - \frac{1}{4} m(m-1)^2(m-2)(p^{-2} - p^{-1}) + \frac{1}{24} m(m-1)(m-2)(3m-1)p^{-2} \right] + O(n^{m-3}) \\
 &= p^m \left[ n^m - \frac{1}{2} m(m-1)n^{m-1} + \frac{1}{24} m(m-1)(m-2)(3m-1)n^{m-2} + \dots \right].
 \end{aligned}$$

The coefficient of  $p^m n^{m-i}$  is thus a polynomial in  $m$  of degree  $i^2$ , which is equal to the coefficient of  $n^{m-i}$  in  $[f(n, m)]^m$ . But this equality is true for all integral values of  $m$  less than  $n$ . It is therefore an identity. That is to say

$\mathcal{E}[\{f(a, m)\}^m] - p^m [f(n, m)]^m$  tends to zero quicker than any negative power of  $n$ . Hence

$$\left[ \frac{f(a, m)}{f(n, m)} \right]^m$$

is an almost unbiased estimate of  $p^m$ . I conjecture that its bias tends to zero with  $(1-p)^n$ .

I have not obtained a rigorous proof that

$$\mathcal{E}[\ln f(a, 0)] = \ln p + \ln f(n, 0)$$

with an error less than any negative power of  $n$ . It is, however, easy to show that

$$\mathcal{E}[\ln f(a, 0)] - \ln f(n, 0) - \ln p$$

tends to zero with  $n^{-4}$  or more rapidly. For from (2)

$$\begin{aligned}
 \ln[f(a, 0)] &= \ln a + \frac{1}{2a} - \frac{1}{12a^2} + \frac{1}{120a^4} + \dots \\
 &= \ln p + \ln n + \ln \left( 1 + \frac{\theta}{pn} \right) + \frac{1}{2pn} \left( 1 + \frac{\theta}{pn} \right)^{-1} - \\
 &\quad - \frac{1}{12p^2 n^2} \left( 1 + \frac{\theta}{pn} \right)^{-2} + \frac{1}{120p^4 n^4} + \dots
 \end{aligned}$$



$$\begin{aligned}
\mathcal{E}[\ln f(a, 0)] &= \ln p + \ln n - \frac{\mathcal{E}(\theta^2)}{2p^2n^2} + \frac{\mathcal{E}(\theta^3)}{3p^3n^3} - \frac{\mathcal{E}(\theta^4)}{4p^4n^4} + \frac{\mathcal{E}(\theta^5)}{5p^5n^5} - \frac{\mathcal{E}(\theta^6)}{6p^6n^6} + \dots \\
&\quad + \frac{1}{2pn} + \frac{\mathcal{E}(\theta^2)}{2p^3n^3} - \frac{\mathcal{E}(\theta^3)}{2p^4n^4} + \frac{\mathcal{E}(\theta^4)}{2p^5n^5} + \dots \\
&\quad - \frac{1}{12p^2n^2} + \frac{\mathcal{E}(\theta^2)}{4p^4n^4} + O(n^{-4}) \\
&= \ln p + \ln n - \frac{(1-p)}{2pn} + \frac{(1-p)(1-2p)}{3p^2n^2} - \frac{3(1-p)^2}{4p^3n^3} - \frac{(1-p)(1-6p+6p^2)}{4p^3n^3} + \\
&\quad + \frac{2(1-p)^2(1-2p)}{p^3n^3} - \frac{5(1-p)^3}{2p^3n^3} + \frac{1}{2pn} + \\
&\quad + \frac{1-p}{2p^2n^2} - \frac{(1-p)(1-2p)}{2p^3n^3} + \frac{3(1-p)^2}{2p^3n^3} - \\
&\quad - \frac{1}{12p^2n^2} - \frac{1-p}{4p^3n^3} + O(n^{-4}) \\
&= \ln p + \ln n + \frac{1}{2n} - \frac{1}{12n^2} + O(n^{-4}) \\
&= \ln p + \ln f(n, 0) + O(n^{-4}).
\end{aligned}$$

Hence we are fully justified in using (5) even when  $a = 1$ , and putting  $\mathcal{E}[\ln f(a, 0) - \ln f(n, 0)] = \ln p$ .

As a curiosity I mention another set of unbiased estimates. The moment-generating function of a binomial distribution is  $(q + pe^t)^n$ . So

$$\begin{aligned}
\mathcal{E}[e^{at}] &= 1 + \sum_{r=1}^{\infty} \frac{t^r}{r!} \mathcal{E}(a^r) \\
&= M(t) \\
&= (q + pe^t)^n.
\end{aligned}$$

Or if  $e^t = k$ , where  $k$  is any positive number

$$\mathcal{E}[k^a] = (q + kp)^n$$

... (8)

thus  $k^a$  is an unbiased estimate of  $(q + kp)^n$ , and in particular  $2^a$  is an unbiased estimate of  $(1+p)^n$ .

Differentiating with regard to  $t$  we find

$$\mathcal{E}[ae^{at}] = npe^t(q + pe^t)^{n-1}$$

or

$$\mathcal{E}[ak^a] = npk(kp + q)^{n-1}.$$

Similarly

$$\mathcal{E}[a^2 k^a] = (n^2 p^2 k^2 + npqk)(kp + q)^{n-2}$$

... (9)

# ALMOST UNBIASED ESTIMATES OF FUNCTIONS OF FREQUENCIES

and a series of similar expressions, which give the values of the moments when  $k = 1$ . I know of no cases where any of these expressions are useful.

I have however used expressions (3), (4) and (5), or rather their first few terms, in several contexts. Expression (3) was used (Haldane, 1956a) in estimating relative viability, and (4) (Haldane, 1956c) in estimating gene frequencies. For if  $p$  is the frequency of a recessive phenotype in a random mating population, then  $p^{\frac{1}{2}}$  is the frequency of the gene concerned. Finally (5) was used in the analysis of fourfold tables (Haldane, 1956b). Consider the table

	$P$	$Q$
$X$	$a$	$b$
$Y$	$c$	$d$

where  $a$  individuals combine the characters  $P$  and  $X$  and so on. For example  $X$  might be inoculation,  $Y$  non-inoculation,  $P$  survival, and  $Q$  death from a disease. Suppose the sample is so large that sampling errors can be neglected. Then if  $\frac{ad}{bc} = e^y$ , Yule's (1912) coefficient of association  $Q = \tanh \frac{1}{2}y$ , while his coefficient of colligation  $Y = \tanh \frac{1}{4}y$ . Now  $e^y = \frac{a}{b} \div \frac{c}{d}$ ; for example in the case of inoculations, the ratio of survivors to dead in the inoculated group divided by the same ratio in the uninoculated group. Woolf (1955) has pointed out that this has a more concrete meaning than the usual coefficients. It is not obvious, for example, that  $Q = \frac{1}{2}$  corresponds to a ratio  $e^y$  of 3.

In a finite sample it is clear that  $\frac{ad}{(b+1)(c+1)}$  is an almost unbiased estimate of  $e^y$ , and

$$\ln f(a, 0) - \ln f(b, 0) - \ln f(c, 0) + \ln f(d, 0)$$

is an almost unbiased estimate of  $y$ . Many other examples could be given, but these should be sufficient to show some of the fields in which unbiased estimates may be useful.

They are unimportant when samples are large. In isolated samples the maximum likelihood estimate is probably preferable. But in series of numerous small samples in which the quantity estimated is believed to differ from one sample to another, unbiased estimates are usually preferable.

Doubtless other functions than powers and logarithms will occasionally be needed. But in the vast majority of cases they can be expressed as power series, and almost unbiased estimates of them can be derived from (3).

Professor C. R. Rao has kindly pointed out to me that the estimates of powers and logarithm of  $p$  here given are not unique. We could clearly add to these estimates any function of  $a$  which tends to zero more rapidly than any negative power of  $n$ , (for example,  $\left(\frac{a}{n}\right)^n$  multiplied by any constant) and the estimate so found would still be almost unbiased in the sense here used. I think, however, that my estimates are the simplest possible.

#### SUMMARY

Estimates of powers of a frequency, and of its logarithm, are obtained whose bias falls off more rapidly than any power of the sample number.

#### REFERENCES

- BHATTACHARYYA, A. (1954): Unbiased and biased statistics in the binomial population. *Cal. Stat. Ass. Bull.*, 5, 149-164.
- DE WINTON, D. AND HALDANE, J. B. S. (1933): The genetics of *Primula sinensis* II. Segregation and interaction of factors in the diploid. *J. Genet.*, 27, 1-44.
- HALDANE, J. B. S. (1933): A class of efficient estimates of a parameter. *Bull. Inter. Stat. Inst.*, 33, 231-248.
- (1956a): The estimation of viabilities. *J. Genet.*, 54 (In press).
- (1956b): The estimation and significance of the logarithm of a ratio of frequencies. *Ann. Hum. Gen.*, 20, 309-311.
- HALDANE, J. B. S. (1956c): Unpublished.
- WOOLF, B. (1955). On estimating the relation between blood group and disease. *Ann. Hum. Gen.*, 19, 251-253.

*Paper received : May, 1956.*

# SUFFICIENT STATISTICS IN ELEMENTARY DISTRIBUTION THEORY

By ROBERT V. HOGG AND ALLEN T. CRAIG

*University of Iowa*

## 1. INTRODUCTION

The following theorem is implicit in papers by Neyman (1936, 1938). In a regular case of estimation, let a probability density function (p.d.f.) in one random variable depend upon a single parameter  $\theta$  that lies in some non-degenerate interval. Let a random sample of  $n$  ( $n > 1$ ) values of the variable afford a single sufficient statistic  $z$  for  $\theta$ . A necessary and sufficient condition that any other statistic  $T$  be stochastically independent of  $z$  is that the distribution of  $T$  be free of  $\theta$ . We extended Neyman's work to a wider class of probability densities, including certain non-regular densities, and one of us applied the theorem (Hogg, 1953, 1956) to problems in non-regular cases. More recently, Basu (1955) has published the result in complete generality. We have found that in its various forms the theorem is a remarkably powerful tool in distribution problems and that, under suitable restrictions, it can be used very effectively by students in a first course in mathematical statistics. With teachers of such courses primarily in mind, we present the following discussion.

## 2. REGULAR CASES

In the interest of conservation of space, we assume that all concepts and definitions needed in the sequel have been introduced. Let  $x_1, x_2, \dots, x_n$  denote  $n$  independent observations of a random variable  $x$  in a one-dimensional Euclidean space which has a p.d.f.  $f(x; \theta_1, \theta_2, \dots, \theta_q)$  depending on  $q$  ( $q < n$ ) unknown parameters  $\theta_1, \theta_2, \dots, \theta_q$  that lie in a non-degenerate  $q$ -dimensional interval. We take  $f(x; \theta_1, \theta_2, \dots, \theta_q)$  to be of the form (Koopman, 1936, Pittman, 1936)

$$Q(\theta_1, \theta_2, \dots, \theta_q)M(x) \exp \left[ \sum_{j=1}^q p_j(\theta_1, \theta_2, \dots, \theta_q) K_j(x) \right], a < x < b, \quad \dots \quad (2.1)$$

- where
- (i)  $M(x)$  is non-negative and continuous,
  - (ii) the  $K'_j(x)$  are continuous and linearly independent,
  - (iii) the  $p_j$  are continuous and the set of values assumed by  $(p_1, p_2, \dots, p_q)$  contains a non-degenerate  $q$ -dimensional interval, and
  - (iv)  $a$  and  $b$  do not depend on the  $\theta_j$ .

For brevity, we call these regular cases and naturally we introduce the student to the subject with  $q = 1$ . The student can verify that  $z_j = \sum_{i=1}^n K_j(x_i)$ ,  $j = 1, 2, \dots, q$  is a set of exactly  $q$  joint sufficient statistics for the  $q$  parameters. By the usual proce-



ture of change of variables, he can show that the simultaneous p.d.f. of the joint sufficient statistics is of the form

$$g(z_1, z_2, \dots, z_q; \theta_1, \theta_2, \dots, \theta_q) \\ = [Q(\theta_1, \theta_2, \dots, \theta_q)]^n R(z_1, z_2, \dots, z_q) \exp \left[ \sum_{j=1}^q p_j(\theta_1, \theta_2, \dots, \theta_q) z_j \right], \alpha_j < z_j < \beta_j,$$

where, under our assumptions,  $R(z_1, z_2, \dots, z_q)$ , the  $\alpha_j$  and the  $\beta_j$  do not depend on the  $\theta_j$ .

We define next, with Lehmann and Scheffé (1950), the notion of completeness of a p.d.f. Let  $\varphi(v_1, \dots, v_m; \theta_1, \dots, \theta_q)$  be a p.d.f. in  $m$  random variables which depends on  $q$  parameters that lie in a non-degenerate  $q$ -dimensional interval. Let  $u(v_1, \dots, v_m)$  be a continuous function of the  $v$ 's (but not a function of the  $\theta$ 's). If

$$\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} u(v_1, \dots, v_m) \varphi(v_1, \dots, v_m; \theta_1, \dots, \theta_q) dv_1 \dots dv_m = 0,$$

for all values of  $\theta_j, j = 1, 2, \dots, q$ , implies that  $u(v_1, \dots, v_m) = 0$ , we say that  $\varphi$  is complete.<sup>1</sup> We then ask the student to accept the non-statistical fact that a certain mathematical concept (the uniqueness of the Laplace bilateral transform) can be used (Lehmann and Scheffé, 1955) to prove under our conditions, that  $g(z_1, \dots, z_q; \theta_1, \dots, \theta_q)$  is always complete.

We consider now any other statistic  $T = T(x_1, \dots, x_n)$  of the random sample. Denote by  $D(T)$  the distribution function of  $T$  and by  $H(T|z_1, \dots, z_q)$  the conditional distribution function of  $T$  given  $z_1, \dots, z_q$ . The student can verify that

$$\int_{\alpha_q}^{\beta_q} \dots \int_{\alpha_1}^{\beta_1} [D(T) - H(T|z_1, \dots, z_q)] g(z_1, \dots, z_q; \theta_1, \dots, \theta_q) dz_1, \dots, dz_q = 0$$

for all values of the  $\theta_j$ .

Since the  $z_j$  are joint sufficient statistics for the  $\theta_j, j = 1, 2, \dots, q$ ,  $H(T|z_1, \dots, z_q)$  is free of the  $\theta_j$  and is, under our conditions, continuous in  $z_j$  for every given value of  $T$ . As soon then as  $D(T)$  is free of the  $\theta_j$ , the completeness of  $g$  implies  $D(T) - H(T|z_1, \dots, z_q) = 0$  so that  $T$  is stochastically independent of  $z_1, \dots, z_q$ . Conversely if  $T$  is stochastically independent of  $z_1, \dots, z_q$  so that

$$D(T) = H(T|z_1, \dots, z_q)$$

we see that  $D(T)$  is free of  $\theta_1, \dots, \theta_q$ . The latter fact of course does not rely on the completeness of  $g$ , but since  $g$  is here always complete, we state the theorem for our students in the form of a necessary and sufficient condition.

**Theorem:** Let  $f(x; \theta_1, \dots, \theta_q)$  denote a p.d.f. of the form (2.1) for which (i)—(iv) are satisfied. Let  $x_1, x_2, \dots, x_n$ , ( $q < n$ ) denote a random sample of  $n$  values of  $x$  so that the  $q$  statistics  $z_j = \sum_{i=1}^n K_j(x_i), j = 1, \dots, q$  are joint sufficient statistics for the  $q$  parameters.

<sup>1</sup> Obviously we take  $u$  to be continuous to avoid discussion of  $u = 0$  (a.e.). That  $\varphi$  is actually strongly complete need not concern the student at this level.

## SUFFICIENT STATISTICS IN ELEMENTARY DISTRIBUTION THEORY

*A necessary and sufficient condition that a statistic  $T$  be stochastically independent of the joint sufficient statistics  $z_j$  is that the distribution of  $T$  be free of the  $\theta_j, j = 1, 2, \dots, q$ .*

Tests of statistical hypotheses rely heavily upon distribution theory. Frequently, the test of a hypothesis will be based upon some statistic  $T$  whose distribution is free of certain nuisance parameters, say  $\theta_1, \theta_2, \dots, \theta_q$ . Thus if, under the null hypothesis, the underlying p.d.f. of  $x$  is of the form (2.1), this statistic  $T$  is stochastically independent of the joint sufficient statistics for the parameters. We now give some examples which illustrate the power and elegance of the theorem. Some of our examples are taken from standard texts but a number of the more interesting are found in the literature.

*Example 2.1:* Let  $f(x; \theta) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(x-\theta)^2\right], -\infty < x < \infty$ .

Here  $z = \sum_{i=1}^n x_i$  is a single sufficient statistic for  $\theta$  so that likewise is  $\bar{x} = z/n$ .

(a) Let  $T = \sum_{i=1}^n (x_i - \bar{x})^k$ . Under the transformation  $y_i = (x_i - \theta)$ , the simultaneous p.d.f. of  $y_1, \dots, y_n$  becomes free of  $\theta$ , and  $\theta$  is not introduced into  $T$ . Hence the characteristic function (or moment generating function) of  $T$  will not contain  $\theta$ , and so the distribution of  $T$  is free of  $\theta$ . Accordingly,  $T$  is stochastically independent of  $\bar{x}$ .

(b) Let  $T = \max(x_i) - \min(x_i)$ . The distribution of  $T$  is free of  $\theta$  so  $T$  is stochastically independent of  $\bar{x}$ .

(c) Let  $T = \sum_{j=1}^n \sum_{i=1}^n a_{ij} x_i x_j, a_{ij} = a_{ji}$ .

With  $y_i = x_i - \theta$ ,

$$T = \sum_{j=1}^n \sum_{i=1}^n a_{ij} y_i y_j + \theta \left[ \sum_{j=1}^n \sum_{i=1}^n a_{ij} y_i + \sum_{j=1}^n \sum_{i=1}^n a_{ij} y_j \right] + \theta^2 \sum_{j=1}^n \sum_{i=1}^n a_{ij}.$$

Thus the distribution of  $T$  will be free of  $\theta$  (and hence  $T$  and  $\bar{x}$  will be stochastically independent) if and only if the sum of the elements of each row of the matrix of  $T$  is zero.

(d) More generally, let  $T = T(x_1, \dots, x_n)$  be any statistic. A necessary and sufficient condition that  $T$  be stochastically independent of  $z$  is that  $T(x_1, \dots, x_n)$  and  $T(x_1 + \theta, \dots, x_n + \theta)$  have identical distributions (Laha, 1956).

*Example 2.2:* Let  $f(x; \theta_1, \theta_2) = \frac{1}{\sqrt{2\pi\theta_2}} \exp\left[-(x-\theta_1)^2/(2\theta_2)\right], -\infty < x < \infty$ .

Here  $z_1 = \sum_{i=1}^n x_i$  and  $z_2 = \sum_{i=1}^n x_i^2$  are joint sufficient statistics for  $\theta_1$  and  $\theta_2$  so likewise

are  $\bar{x} = z_1/n$  and  $s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$ .

(a) Vide von Neumann (1941)

$$\text{Let } T = \frac{\sum_{i=1}^{n-1} (x_{i+1} - x_i)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

The transformation  $y_i = \frac{(x_i - \theta_1)}{\sqrt{\theta_2}}$  shows that the distribution of  $T$  is free of  $\theta_1$  and  $\theta_2$ .

Thus  $T$  is stochastically independent of both  $\bar{x}$  and  $s^2$ . Two similar examples may be found in Madow (1945) and Lehmann (1947).

(b) Let  $T = [\max(x_i) - \bar{x}] / [\max(x_i) - \min(x_i)]$  (McKay, 1935). The distribution of  $T$  is free of  $\theta_1$  and  $\theta_2$  so that  $T$  is stochastically independent of both  $\bar{x}$  and  $s^2$ .

*Example 2.3:* Let  $f(x; \theta) = \left( 1 / [\Gamma(\alpha + 1) \theta^{\alpha+1}] \right) x^\alpha \exp[-x/\theta]$ ,  $0 < x < \infty$ .

Again  $z = \sum_{i=1}^n x_i$  is a single sufficient statistic for  $\theta$ . Let  $T = T(x_1, \dots, x_n)$  be any statistic. A necessary and sufficient condition that  $T$  and  $z$  be stochastically independent is that  $T(x_1, \dots, x_n)$  have a distribution identical with  $T(x_1/\theta, \dots, x_n/\theta)$ ; for evidently the latter has a distribution free of  $\theta$ . (Laha, 1956).

### 3. OTHER REGULAR CASES

In section 2 we restricted ourselves to a probability density  $f$  in one random variable  $x$  and to a random sample of  $n > 1$  values of that variable. Suppose now we let  $f$  denote the simultaneous p.d.f. of a random sample point  $e$  with dimension of  $e$  greater than  $q$  and take  $f$  to be of the form

$$f(e; \theta_1, \dots, \theta_q) = Q_1(\theta_1, \dots, \theta_q) M_1(e) \exp \left[ \sum_{j=1}^q z_j(e) p_j(\theta_1, \dots, \theta_q) \right]$$

where (i) the  $z_j$  are functionally independent and have continuous partial derivatives with respect to the items of the sample,

(ii) the  $p_j$  are continuous and the set of values assumed by  $(p_1, \dots, p_q)$  contains a non-degenerate  $q$ -dimensional interval,

and (iii) the domain of  $f$  does not depend upon the  $q$  parameters.

Then the simultaneous p.d.f. of the joint sufficient statistics  $z_1, \dots, z_q$  is complete, (Lehmann and Scheffé, 1955) and the conclusion of the theorem of § 2 is valid.

*Example 3.1:* Let  $(x_i, y_i)$ ,  $i = 1, 2, \dots, n$ , be a random sample of  $n$  values of  $x$  and  $y$  from a normal bivariate distribution. If  $\rho$ , the correlation coefficient between  $x$  and  $y$ , is zero, the two sample means,  $\bar{x}, \bar{y}$ , and the two sample variances,  $s_x^2, s_y^2$ , are four mutually stochastically independent joint sufficient statistics for the two means,  $\mu_x, \mu_y$ , and the two variances,  $\sigma_x^2, \sigma_y^2$ . Since, when  $\rho = 0$ , the distribution of the sample correlation coefficient  $r$  does not depend upon these four parameters,  $r$  is stochastically independent of the joint sufficient statistics. Thus, if momentarily, we set  $\rho = \mu_x = \mu_y = 0$ ,  $\sigma_x^2 = \sigma_y^2 = 1$ , the simultaneous p.d.f.,



## SUFFICIENT STATISTICS IN ELEMENTARY DISTRIBUTION THEORY

say  $g_0$ , of  $\bar{x}$ ,  $\bar{y}$ ,  $s_x^2$ ,  $s_y^2$ , and  $r$ , is the product of the five particularly simple marginal densities. We may now proceed to find the simultaneous p.d.f., say  $g$ , of the five joint sufficient statistics for all admissible values of the five parameters by using a theorem of Madow (1938); that is,

$$g = (g_0) \left( \frac{L(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)}{L(0, 0, 1, 1, 0)} \right)$$

where  $L(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$  is the joint p.d.f. of the sample items.

*Example 3.2:*<sup>1</sup> Let  $x_1, \dots, x_n$  and  $y_1, \dots, y_m$  be random samples from two independent normal distributions. The statistic is used to test

$$F = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 / (n-1)}{\sum_{i=1}^m (y_i - \bar{y})^2 / (m-1)}$$

the hypothesis  $H$  that the population variances are equal. The statistics  $z_1 = \bar{x}$ ,  $z_2 = \bar{y}$  and  $z_3 = \sum_{i=1}^n (x_i - \bar{x})^2 + \sum_{i=1}^m (y_i - \bar{y})^2$  are joint sufficient statistics for the three parameters, namely the two means and the common variance. Since, under  $H$ , the distribution of  $F$  is free of these parameters,  $F$  and the function of the joint sufficient statistics

$$t = \frac{\left[ \frac{nm}{n+m} \right]^{\frac{1}{2}} (\bar{x} - \bar{y})}{\left[ \frac{\sum_{i=1}^n (x_i - \bar{x})^2 + \sum_{i=1}^m (y_i - \bar{y})^2}{n+m-2} \right]^{\frac{1}{2}}}$$

are stochastically independent. Thus, assuming equal variances,  $F$  is stochastically independent of  $t$  irrespective as to whether  $t$  is or is not central.

### 4. NON-REGULAR CASES

In this section we consider some non-regular cases that would normally not be presented in an elementary course, but could be used as special exercises for the better students. We let  $x_1, \dots, x_n$  denote  $n$  ordered, initially independent observations of a random variable  $x$  in a one-dimensional Euclidean space which has a p.d.f.  $f(x; \theta)$  depending on one unknown parameter  $\theta$  that lies in a non-degenerate interval. We take  $f(x; \theta)$  to be of the form

$$Q(\theta)M(x), \quad a(\theta) < x < b(\theta), \quad \dots \quad (4.1)$$

where (i)  $M(x)$  is positive and continuous,

(ii)  $a'(\theta)$  and  $b'(\theta)$  are continuous, and

either (iii)  $a(\theta)$  is constant,  $b(\theta)$  strictly monotone (or vice versa) and  $\sup a(\theta) = \inf b(\theta)$ ,

or (iv)  $a(\theta)$  is strictly monotone decreasing,  $b(\theta)$  strictly monotone increasing (or vice versa) and  $\sup a(\theta) = \inf b(\theta)$ .

Under these conditions it is known (Pitman, 1936 ; Davis, 1951), there exists a single sufficient statistic for the parameter.

<sup>1</sup> Suggested by R. R. Bahadur.



Under (iii) the single sufficient statistic is  $z = x_n$  (or  $z^* = x_1$ ). The p.d.f. of  $z$  is

$$g(z; \theta) = n[Q(\theta)]^n \left[ \int_{a(\theta)}^z M(x) dx \right]^{n-1} M(z) \quad a(\theta) < z < b(\theta),$$

and may be shown to be complete as follows. Let  $u(z)$  be continuous and set

$$\int_{a(\theta)}^{b(\theta)} u(z) g(z; \theta) dz = 0$$

for all  $\theta$ .

Thus,

$$\int_{a(\theta)}^{b(\theta)} u(z) \left[ \int_a^z M(x) dx \right]^{n-1} M(z) dz = 0$$

for all  $\theta$ .

If we differentiate both members with respect to  $b(\theta)$ , we find that  $u(b(\theta)) \left[ \int_{a(\theta)}^{b(\theta)} M(x) dx \right]^{n-1} M(b(\theta)) = 0$  for all  $\theta$ ; whence  $u(z) = 0$  for  $a(\theta) < z < b(\theta)$ .

Under condition (iv), the single sufficient statistic is  $z = \max [a^{-1}(x_1), b^{-1}(x_n)]$  {or  $z^* = \min [a^{-1}(x_1), b^{-1}(x_n)]$ }. The distribution function of  $z$  is

$$\begin{aligned} G(z; \theta) &= P[a^{-1}(x_1) \leq z, b^{-1}(x_n) \leq z] \\ &= P[a(z) \leq x_1, x_n \leq b(z)] \\ &= \left[ \int_{a(z)}^{b(z)} Q(\theta) M(x) dx \right]^n. \end{aligned}$$

Hence the p.d.f. of  $z$ ,  $g(z; \theta)$ , is

$$n[Q(\theta)]^n \left[ \int_{a(z)}^{b(z)} M(x) dx \right]^{n-1} \left[ M(b(z)) \frac{db}{dz} - M(a(z)) \frac{da}{dz} \right], \quad c < z < \theta,$$

where  $c = \inf [\theta; a(\theta) < b(\theta)]$ . We show  $g(z; \theta)$  to be complete as before. Accordingly, the conclusion of the theorem of section 2 is valid when sampling from probability densities of the form (4.1).

**Example 4.1:** (Darling 1952). Let  $x_1, \dots, x_n$  be an ordered sample from a uniform distribution over  $(0, \theta)$ . Since  $T = (x_1 + x_2 + \dots + x_n)/x_n$  is distributed free of  $\theta$ ,  $T$  and the sufficient statistic  $z = x_n$  are stochastically independent. Thus  $D(T) = H(T|z)$ . In particular, let  $z = 1$ . Then  $T$  is distributed like  $1 + y_1 + \dots + y_{n-1}$ , where  $y_1, \dots, y_{n-1}$  are independent and are uniformly distributed over  $(0, 1)$ .

**Example 4.2:** (Epstein and Sobel, 1954). Let

$$f(x; \theta, \rho) = (1/\rho) \exp [-(x-\theta)/\rho], \quad \theta < x < \infty.$$

The smallest item,  $x_1$ , of an ordered sample of size  $n$  is, for every fixed  $\rho$ , a sufficient statistic for  $\theta$ . Hence  $x_1$  is stochastically independent of every statistic whose distribution is free of the location parameter  $\theta$ ; in particular,  $x_1$  is stochastically independent of the statistic  $T = \sum_{i=1}^r (x_i - x_1) + (n-r)(x_r - x_1)$ ,  $r < n$ .

## SUFFICIENT STATISTICS IN ELEMENTARY DISTRIBUTION THEORY

*Example 4.3:* Let  $x_1, \dots, x_n$  be an ordered sample from a uniform distribution over  $(\mu - \theta, \mu + \theta)$ . The likelihood ratio used to test the hypothesis  $\mu = 0$  is  $\lambda = [(x_n - x_1)/2z]^n$  where  $z = \max(-x_1, x_n)$  is the sufficient statistic for the nuisance parameter  $\theta$ . Since, under the null hypothesis, the distribution of  $\lambda$  is free of  $\theta$ , the theorem asserts the stochastic independence of  $z$  and  $\lambda$ . This permits us to write

$$E[\exp\{it(-2\ln \lambda)\}] = \frac{E[\exp\{it[-2\ln(x_n - x_1)^n]\}]}{E[\exp\{it[-2\ln(2z)^n]\}]}.$$

We now set  $\theta = \frac{1}{2}$ , take the limit of each member of the above equation as  $n$  becomes infinite, and show, after some simplification, that  $-2\ln \lambda$  has a limiting  $\chi^2$ -distribution with two degrees of freedom.

In from (4.1) we restricted ourselves to a p.d.f. which depends on one parameter. We now introduce two parameters and we let the end-points of the range be functionally independent by considering probability densities of the form

$$Q(\theta_1, \theta_2)M(x), \quad a(\theta_1, \theta_2) < x < b(\theta_1, \theta_2), \quad \dots \quad (4.2)$$

where (i)  $M(x)$  is positive and continuous,

and (ii)  $a(\theta_1, \theta_2)$  and  $b(\theta_1, \theta_2)$  are monotone and continuous in each parameter and the values assumed by  $(a, b)$  is a non-degenerate two-dimensional set of which one boundary is  $a = b$ .

As before, it is easy to show that the joint sufficient statistics  $z_1 = x_1$  and  $z_2 = x_n$  have a complete simultaneous p.d.f. by partial differentiation with respect to the end-points of the range. Again the conclusion of the theorem of § 2 is valid.

It is interesting to observe that the underlying p.d.f. of  $x$  can be a mixture of the regular and non-regular types and still have the simultaneous p.d.f. of the joint sufficient statistics complete. This is the case, for example, if the p.d.f of  $x$  is of the form

$$Q(\theta_1, \theta_2, \dots, \theta_q)M(x) \exp \left[ \sum_{j=3}^q p_j(\theta_3, \dots, \theta_q) K_j(x) \right], \quad a(\theta_1, \theta_2) < x < b(\theta_1, \theta_2), \quad \dots \quad (4.3)$$

where conditions (i) and (ii) of (4.2) and conditions (i)-(iv) of (2.1) (with obvious modifications) are satisfied. Here the  $q$  joint sufficient statistics, whose simultaneous p.d.f. is complete, are  $z_1 = x_1$ ,  $z_2 = x_n$  and  $z_j = \sum_{i=1}^n K_j(x_i)$ ,  $j = 3, 4, \dots, q$ . It is important to note, however, that here, as previously, the number of joint sufficient statistics is equal to the number of parameters.

*Example 4.4:* (Hogg 1953). Let  $x_1, \dots, x_n$  and  $y_1, \dots, y_m$  be ordered samples from two rectangular populations having equal but unknown ranges. We use  $t = \max(x_n - x_1, y_m - y_1) / [\max(x_n, y_m) - \min(x_1, y_1)]$  to test the hypothesis that the two means are equal. Under this hypothesis,  $z_1 = \min(x_1, y_1)$  and  $z_2 = \max(x_n, y_m)$  are joint sufficient statistics for the unknown end-points of the distributions. Since  $t$  has a distribution free of these parameters,  $t$  and  $(z_1, z_2)$  are stochastically

independent. Hence the ratio  $t$  and its denominator are stochastically independent so that the moments of  $t$  are equal to the ratios of the corresponding moments of its numerator and denominator.

*Example 4.5 :* (Paulson, 1941). Let

$$f(x; \theta_1, \theta_2) = (1/\theta_2) \exp [-(x-\theta_1)/\theta_2], \quad \theta_1 < x < \infty.$$

If  $x_1, \dots, x_n$  and  $y_1, \dots, y_m$  are two ordered samples from this population, then  $z_1 = \min(x_1, y_1)$  and  $z_2 = \sum_{i=1}^n x_i + \sum_{i=1}^m y_i$  are joint sufficient statistics for the parameters. Thus

$$\lambda = \left[ \frac{\sum_{i=1}^n (x_i - x_1) + \sum_{i=1}^m (y_i - y_1)}{z_2 - (n+m)z_1} \right]^{n+m},$$

which is distributed free of  $\theta_1$  and  $\theta_2$ , is stochastically independent of  $z_1$  and  $z_2$  and hence of its denominator.

#### REFERENCES

- BASU, D. (1955): On statistics independent of a complete sufficient statistic. *Sankhyā*, **15**, 377-380.
- DARLING, D. A. (1952): On a test for homogeneity and extreme values. *Ann. Math. Stat.*, **23**, 450-456.
- DAVIS, R. C. (1951): On minimum variance in non-regular estimation. *Ann. Math. Stat.*, **22**, 43-57.
- EPSTEIN, B. AND SOBEL, M. (1954): Some theorems relevant to life testing from an exponential distribution. *Ann. Math. Stat.*, **25**, 373-381.
- HOGG, R. V. (1953): Testing the equality of means of rectangular populations. (Abstract). *Ann. Math. Stat.*, **24**, 691.
- (1956): On the distribution of the likelihood ratio. *Ann. Math. Stat.*, **27**, 529.
- KOOPMAN, B. O. (1936): On distributions admitting a sufficient statistic. *Trans. Amer. Math. Soc.*, **39**, 399-409.
- LAHA, R. G. (1956): On some properties of the normal and gamma distributions. *Proc. Amer. Math. Soc.*, **7**, 172-174.
- LEHMANN, E. L. (1947): On optimum tests of composite hypotheses with one constraint. *Ann. Math. Stat.*, **18**, 473-494.
- LEHMANN, E. L. AND SCHEFFÉ, HENRY (1950): Completeness, similar regions, and unbiased estimation. *Sankhyā*, **10**, 305-340.
- (1955): Completeness, similar regions, and unbiased estimation. *Sankhyā*, **15**, 219-236.
- MADOW, W. G. (1938): Contributions to the theory of multivariate statistical analysis. *Trans. Amer. Math. Soc.*, **44**, 454-495.
- (1945): Note on the distribution of the serial correlation coefficient. *Ann. Math. Stat.*, **16**, 308-310.
- McKAY, A. T. (1935): The distribution of the difference between the extreme observation and the sample mean in samples of  $n$  from a normal universe. *Biometrika*, **27**, 466-471.
- NEYMAN, J. (1936): Sur la vérification des hypothèses statistiques composées. *Bull. Soc. Math. Franch.*, **39**, 1-21.
- (1938): On statistics the distribution of which is independent of the parameters involved in the original probability law of the observed variables. *Stat. Res. Mem.*, **2**, 58-59.
- NEUMANN, J. VON (1941): Distribution of the ratio of the mean square successive difference to the variance. *Ann. Math. Stat.*, **12**, 367-395.
- PAULSON, EDWARD (1941): On certain likelihood ratio tests associated with the exponential distribution. *Ann. Math. Stat.*, **12**, 301-306.
- PITTMAN, E. J. G. (1936): Sufficient statistics and intrinsic accuracy. *Proc. Camb. Phil. Soc.*, **32**, 567-579.

*Paper received : August, 1956.*



# MISCELLANEOUS

## SUFFICIENT STATISTICS AND ORTHOGONAL PARAMETERS

By V. S. HUZURBAZAR

*University of Poona*

### 1. INTRODUCTION

Let  $f(x, \alpha_i)$  be the probability density function of a distribution depending on  $n$  parameters  $\alpha_i$  ( $i = 1, 2, \dots, n$ ). Then the parameters  $\alpha_i$  are said to be orthogonal (Jeffreys, 1948) if

$$E_{ij} = E \left\{ - \frac{\partial^2}{\partial \alpha_i \partial \alpha_j} \log f \right\} \equiv 0, \quad \text{for all } i, j \ (i \neq j),$$

In general, we have a choice in choosing the form of the parameters of a distribution. If the  $\beta_j$  are any  $n$  independent functions of the  $\alpha_i$ 's, we may take the  $\beta_j$  as parameters in place of the  $\alpha_i$ 's. The importance of orthogonal parameters lies in the fact that their maximum likelihood estimators are uncorrelated, and hence a practical solution is obtained with great ease by the method of iteration.

The general problem of finding orthogonal parameters for a probability distribution, has been investigated by the present author (Huzurbazar, 1950) who has pointed out the difficulties and complexities inherent in the problem, and considered some of the cases in which the problem is solvable.

In this note, it is shown that when a probability distribution depending on two parameters admits jointly sufficient statistics, the problem of finding orthogonal parameters is solvable.

### 2. TWO-PARAMETER DISTRIBUTIONS ADMITTING SUFFICIENT STATISTICS

Following Koopman (1936) the most general form of such distributions is

$$f(x, \alpha_1, \alpha_2) = \exp \{u_1(\alpha_1, \alpha_2) v_1(x) + u_2(\alpha_1, \alpha_2) v_2(x) + A(x) + B(\alpha_1, \alpha_2)\} \quad \dots \quad (1)$$



where  $u_1, u_2, B$  are functions of  $\alpha_1, \alpha_2$ ; and  $v_1, v_2, A$  are functions of  $x$ .

We may take  $u_1 = u_1(\alpha_1, \alpha_2)$  and  $u_2 = u_2(\alpha_1, \alpha_2)$  as the new parameters and express  $B$  in terms of  $u_1$  and  $u_2$ .

Then we can write

$$f(x, u_1, u_2) = \exp \{u_1 v_1(x) + u_2 v_2(x) + A(x) + B(u_1, u_2)\}. \quad \dots (2)$$

We have

$$\left. \begin{aligned} E_{11} &= E \left\{ -\frac{\partial^2}{\partial u_1^2} \log f \right\} = -\frac{\partial^2 B}{\partial u_1^2} \\ E_{12} &= E \left\{ -\frac{\partial^2}{\partial u_1 \partial u_2} \log f \right\} = -\frac{\partial^2 B}{\partial u_1 \partial u_2} \\ E_{22} &= E \left\{ -\frac{\partial^2}{\partial u_2^2} \log f \right\} = -\frac{\partial^2 B}{\partial u_2^2} \end{aligned} \right\}. \quad \dots (3)$$

Now consider the transformation

$$\left. \begin{aligned} u_1 &= \beta_1 \\ u_2 &= u_2(\beta_1, \beta_2) \end{aligned} \right\} \quad \dots (4)$$

where we try to adjust  $u_2$  so that the parameters  $\beta_1$  and  $\beta_2$  are orthogonal. To do this, we have to solve the differential equation (Huzurbazar, 1950),

$$E_{12} + E_{22} \frac{\partial u_2}{\partial \beta_1} = 0. \quad \dots (5)$$

Substituting from (3), and putting  $u_1 = \beta_1$ , (5) becomes

$$\frac{\partial^2 B}{\partial \beta_1 \partial u_2} + \frac{\partial^2 B}{\partial u_2^2} \cdot \frac{\partial u_2}{\partial \beta_1} = 0. \quad \dots (6)$$

Since  $B$  is a function of  $\beta_1$  and  $u_2$ , where  $u_2$  is itself a function of  $\beta_1$  and  $\beta_2$ , (6) can be written as

$$\frac{\partial}{\partial \beta_1} \left\{ \frac{\partial B}{\partial u_2} \right\} = 0, \quad \dots (7)$$

which gives

$$\frac{\partial B}{\partial u_2} = \psi(\beta_2), \quad \dots (8)$$

# SUFFICIENT STATISTICS AND ORTHOGONAL PARAMETERS

where  $\psi(\beta_2)$  is an arbitrary function of  $\beta_2$ .

Taking for simplicity  $\psi(\beta_2) = \beta_2$ , we have thus shown that the parameters  $\beta_1$  and  $\beta_2$  are orthogonal, where  $\beta_1$  and  $\beta_2$  are defined by

$$\left. \begin{aligned} \beta_1 &= u_1 \\ \beta_2 &= \frac{\partial B}{\partial u_2} \end{aligned} \right\} \dots (9)$$

In like manner, it follows that the transformation

$$\left. \begin{aligned} \beta_1 &= \frac{\partial B}{\partial u_1} \\ \beta_2 &= u_2 \end{aligned} \right\} \dots (10)$$

also yields  $\beta_1$  and  $\beta_2$  as orthogonal parameters.

## 3. EXAMPLE

Consider the two-parameter Type III distribution

$$f(x, a, p) = \frac{a^p e^{-ax} x^{p-1}}{\Gamma(p)} \quad (0 \leq x < \infty).$$

We have

$$f(x, a, p) = \exp \{-ax + p \log x - \log x + p \log a - \log \Gamma(p)\}.$$

Here

$$u_1 = a, u_2 = p, B = p \log a - \log \Gamma(p)$$

$$B(u_1, u_2) = u_2 \log u_1 - \log \Gamma(u_2)$$

$$\frac{\partial B}{\partial u_1} = \frac{u_2}{u_1}.$$

Hence the transformation

$$\left. \begin{aligned} \beta_1 &= \frac{u_2}{u_1} = \frac{p}{a} \\ \beta_2 &= u_2 = p \end{aligned} \right\}$$

yields  $\beta_1$  and  $\beta_2$  as orthogonal parameters.

It will be interesting to know whether the above result can be extended to distributions with three or more parameters admitting sufficient statistics. As pointed out by the present author, the problem becomes exceedingly difficult for the case of three parameters, and in general impossible (except in special cases) when the number of parameters exceeds three.

REFERENCES

- HUZUBAZAR, V. S. (1950) : Probability distributions and orthogonal parameters. *Proc. Camb. Phil. Soc.* 46, 281.
- JEFFREYS, H. (1948) : *Theory of Probability*, Second Edition, Oxford.
- KOOPMAN, B. O. (1936) : On distributions admitting a sufficient statistic. *Trans. Amer. Math. Soc.*, 39, 399.

*Paper received : March, 1955.*

# A NOTE ON THE MULTIVARIATE EXTENSION OF SOME THEOREMS RELATED TO THE UNIVARIATE NORMAL DISTRIBUTION

By D. BASU

*Indian Statistical Institute, Calcutta*

## 1. INTRODUCTION

This note is of an expository nature. It is hoped that the approach will be found useful to the students. It is shown how certain theorems connected with the univariate normal ( $N_1$ ) distribution may be immediately generalized to the multivariate case if we define the multivariate normal distribution as follows (Frechet, 1951).

*Definition:* The random  $p$ -vector  $\mathbf{x}$  is said to have the  $p$ -variate normal ( $N_p$ ) distribution if for every constant  $p$ -vector  $\mathbf{t}$  the distribution of  $\mathbf{t}\mathbf{x}'$  is univariate normal ( $N_1$ ).

That the above definition is equivalent to the usual definition is proved as follows. Since every linear function (functional) of  $\mathbf{x}$  is  $N_1$ , the dispersion matrix  $\Lambda$  of  $\mathbf{x}$  must exist. Let  $\mu$  be the mean vector of  $\mathbf{x}$ . Then, for every  $p$ -vector  $\mathbf{t}$ , the distribution of  $\mathbf{t}\mathbf{x}'$  is  $N_1$  with mean  $\mathbf{t}\mu'$  and variance  $\mathbf{t}\Lambda\mathbf{t}'$

$$E e^{i\mathbf{t}\mathbf{x}'} = e^{i\mathbf{t}\mu' - \frac{1}{2}\mathbf{t}\Lambda\mathbf{t}'}$$

Hence

and the rest follows (Cramer 1946).

## 2. SOME PROPERTIES OF $N_p$

**Theorem 1:** If  $\mathbf{x}$  is  $N_p$  and  $A$  is any constant  $p \times q$  matrix then  $\mathbf{x}A$  is  $N_q$ .

*Proof:* For any  $q$ -vector  $\mathbf{t}$

$$\mathbf{t}(\mathbf{x}A)' = (\mathbf{t}A')\mathbf{x}'$$

But  $(\mathbf{t}A')\mathbf{x}'$  is  $N_1$  because  $\mathbf{x}$  is  $N_p$ .

**Theorem 2:** If  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$  are mutually independent  $N_p$ 's then for any set of constants  $c_1, c_2, \dots, c_n$  the  $p$ -vector  $\mathbf{y} = c_1\mathbf{x}_1 + \dots + c_n\mathbf{x}_n$  is  $N_p$ .

*Proof:* For any  $p$ -vector  $\mathbf{t}$ , the set of random variables  $\mathbf{t}\mathbf{x}_1', \mathbf{t}\mathbf{x}_2', \dots, \mathbf{t}\mathbf{x}_n'$  are mutually independent  $N_1$ 's and hence

$$\mathbf{t}\mathbf{y}' = \sum c_j \mathbf{t}\mathbf{x}_j' \text{ is } N_1$$

**Theorem 3:** If  $\mathbf{x}_1$  and  $\mathbf{x}_2$  are independent random  $p$ -vectors and  $\mathbf{x}_1 + \mathbf{x}_2$  is  $N_p$  then both  $\mathbf{x}_1$  and  $\mathbf{x}_2$  are  $N_p$ 's (a constant  $p$ -vector is a degenerate case of  $N_p$ .)



*Proof:* Since  $\mathbf{x}_1 + \mathbf{x}_2$  is  $N_p$  we have, for every  $\mathbf{t}$ ,

$$\mathbf{t}\mathbf{x}_1' + \mathbf{t}\mathbf{x}_2' = \mathbf{t}(\mathbf{x}_1 + \mathbf{x}_2)' \text{ is } N_1.$$

And since  $\mathbf{t}\mathbf{x}_1'$  and  $\mathbf{t}\mathbf{x}_2'$  are independent it follows (Cramer, 1937) that they are both  $N_1$ 's.

**Theorem 4:** If  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$  are mutually independent  $N_p$ 's with common dispersion matrix  $\Lambda$  and

$$\mathbf{y}_i = \sum_j a_{ij} \mathbf{x}_j \quad i, j = 1, 2, \dots, n$$

where  $A = (a_{ij})$  is a unitary orthogonal matrix then  $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n$  are mutually independent  $N_p$ 's with common dispersion matrix  $\Lambda$ .

*Proof:* Let  $\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_n$  be arbitrary  $p$ -vectors and let  $\mathbf{z} = (z_1, z_2, \dots, z_n)$  where

$$z_i = \mathbf{t}_i \mathbf{y}_i' = \sum_j a_{ij} (\mathbf{t}_i \mathbf{x}_j') \quad i = 1, 2, \dots, n.$$

Now, every linear functional of  $\mathbf{z}$  is a linear combination of linear functionals of  $\mathbf{x}_j$ 's and hence is  $N_1$ . That is  $\mathbf{z}$  is  $N_p$ .

It is easily verified that

$$\text{cov}(\mathbf{z}_i \mathbf{z}_s) = \begin{cases} 0 & \text{if } i \neq s \\ \mathbf{t}_i \Lambda \mathbf{t}_i' & \text{if } i = s \end{cases}$$

Hence  $z_i = \mathbf{t}_i \mathbf{y}_i'$ ,  $i = 1, 2, \dots, n$ , are mutually independent normal variables and since this is true whatever be the  $\mathbf{t}_i$ 's it follows that  $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n$  are mutually independent normal variables (To prove this just consider the joint characteristic function of  $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n$ ). That the dispersion matrix of  $\mathbf{y}_i$ ,  $i = 1, 2, \dots, n$ , is  $\Lambda$  follows easily from the fact that the variance of  $\mathbf{t}_i \mathbf{y}_i'$  is  $\mathbf{t}_i \Lambda \mathbf{t}_i'$  for all  $\mathbf{t}_i$ .

**Theorem 5:** If  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$  are mutually independent random  $p$ -vectors such that for two given sets of constants  $a_1, a_2, \dots, a_n$  and  $b_1, b_2, \dots, b_n$  the random  $p$ -vectors

$$\mathbf{y}_1 = \sum a_i \mathbf{x}_i \text{ and } \mathbf{y}_2 = \sum b_i \mathbf{x}_i$$

are independent then every  $\mathbf{x}_i$ , for which  $a_i b_i \neq 0$ , must be an  $N_p$  ( $i = 1, 2, \dots, n$ ).

This is the multivariate extension of the corresponding well-known result for  $p = 1$ . (Basu, 1951; Darmois, 1953; Skitovitch, 1953).

*Proof:* For any  $p$ -vector  $\mathbf{t}$ , the random variable

$$\mathbf{t} \mathbf{y}_1' = \sum a_i (\mathbf{t} \mathbf{x}_i')$$

is independent of

$$\mathbf{t} \mathbf{y}_2' = \sum b_i (\mathbf{t} \mathbf{x}_i')$$

and hence  $\mathbf{t} \mathbf{x}_i'$  is  $N_1$  if  $a_i b_i \neq 0$ .

# A NOTE ON THE MULTIVARIATE EXTENSION OF SOME THEOREMS

Geary (1936) proved under certain restrictive assumptions that if, for  $n$  independent observations  $x_1, x_2, \dots, x_n$  on the random variable  $x$ , the sample mean is independent of the sample variance then  $x$  must be normal. Extending Geary's result Laha (1953) proved that if  $x$  has finite variance  $\sigma^2$  and if there exists an unbiased (whatever be the distribution of  $x$ ) quadratic estimator  $\mathbf{x} A \mathbf{x}'$  of  $\sigma^2$  with the property that

$$E[\mathbf{x} A \mathbf{x}' \mid \Sigma x_i] \equiv \sigma^2$$

then  $x$  must be normal. The multivariate extension (Lukacs, 1942 ; Laha, 1955) of the above result is what follows.

Let  $\mathbf{x} = (x_1, x_2, \dots, x_p)$  be a random  $p$ -vector with finite dispersion matrix  $\Lambda$  and let

$$\mathbf{x}_j = (x_{1j}, x_{2j}, \dots, x_{pj}) \quad j = 1, 2, \dots, n$$

be  $n$  independent observations on  $\mathbf{x}$ .

Let  $X = (x_{ij}) \quad i = 1, 2, \dots, p; \quad j = 1, 2, \dots, n$

be the matrix of observations and let  $XAX'$  be an unbiased (whatever be the distribution of  $\mathbf{x}$ ) estimator of  $\Lambda$ . Let  $\mathbf{s} = (s_1, s_2, \dots, s_p)$  where  $s_i = \sum_j x_{ij}$

Theorem 6 : If  $E[XAX' \mid \mathbf{s}] \equiv \Lambda$  then  $\mathbf{x}$  must be  $N_p$ .

*Proof:* Let  $\mathbf{t}$  be an arbitrary  $p$ -vector.

Note that  $\mathbf{t}X = (\mathbf{t} \mathbf{x}_1', \mathbf{t} \mathbf{x}_2', \dots, \mathbf{t} \mathbf{x}_n')$  is the vector of  $n$  independent observations on  $\mathbf{t} \mathbf{x}'$  and that

$$V(\mathbf{t} \mathbf{x}') = \mathbf{t} \Lambda \mathbf{t}'. \quad \text{Also } \Sigma \mathbf{t} \mathbf{x}_j' = \mathbf{t} \mathbf{s}'.$$

Since  $E(XAX') = \Lambda$

we have  $E[(\mathbf{t}X)A(\mathbf{t}X)'] = \mathbf{t}(E XAX')\mathbf{t}' = \mathbf{t} \Lambda \mathbf{t}'.$

Again since  $E(XAX' \mid \mathbf{s}) \equiv \Lambda$  we have

$$E[(\mathbf{t}X)A(\mathbf{t}X)' \mid \mathbf{t} \mathbf{s}'] \equiv \mathbf{t} \Lambda \mathbf{t}'.$$

Thus, all the requirements of Laha's extension of Geary's theorem are satisfied. Hence  $\mathbf{t} \mathbf{x}'$  is  $N_1$ . Since  $\mathbf{t}$  is arbitrary,  $\mathbf{x}$  must be  $N_p$ .

## REFERENCES

- BASU, D. (1951): On the independence of linear functions of independent chance variables. *Bull. Int. Stat. Inst.*, **33**, Part II, 83-96.
- CRAMER, H. (1936): Über eine Eigenschaft der normalen Verteilungsfunktion. *Math. Zeitschrift*, **41**, 405-414.
- (1937): *Random Variables and Probability distributions*. Cambridge Tracts, No. 36. Cambridge University Press, 52.
- (1946): *Mathematical Methods of Statistics*, Princeton University Press, 310.
- DARMOIS, G. (1953): Analyse generale des liaisons stochastiques etude particuliere de l'analyse factorielle lineaire, *Rev. Inst. Int. Stat.*, **21**, 2-8.
- FRECHET, M. (1951): Generalisation de la loi de probabilite' de Laplace. *Annales de l'Institute Henri Poincare*, **12**, Fasc. L.
- GEARY, R. C. (1936): The distribution of student's ratio for non-normal samples. *J. Roy. Stat. Soc. Suppl.*, **3**, 178-184.
- LAHA, R. G. (1953): On an extension of Geary's theorem. *Biometrika*, **40**, 228-229.
- (1955): On a characterisation of the multivariate normal distribution. *Sankhya*, **14**, 367-368.
- LURACS, L. (1942): A characterisation of the normal distribution. *Ann. Math. Stat.*, **13**, 91-93.
- SKITOVITCH (1953): *On a property of the normal law* (in Russian), Doklady Acad. Science, U.S.S.R.

*Paper received : May, 1955.*

# THE SURVEY RESEARCH CENTRE OF THE CENTRAL BUREAU OF STATISTICS, SWEDEN

By TORE DALENIUS

*Central Bureau of Statistics, Stockholm<sup>1</sup>*

## 1. ORIGIN OF THE SURVEY RESEARCH CENTRE

The radical transformation undergone by Swedish society, changing it into a modern industrial community instead of being one formed largely of peasants, has naturally lent its mark in the official production of statistical data in Sweden as regards content, comprehensiveness, organisation and methods used. In these developments, the use of sampling methods plays an increasingly important part, especially from the year 1945 and onwards.

It is particularly interesting to note that interview sample surveys now-a-days are used to a much greater extent than was formerly the case in the official production of statistical data in Sweden. The conditions required for them were not favourable in many respects; for instance, the production of statistical data in Sweden is highly de-centralised, which made it difficult to carry out investigations requiring a nation-wide field organisation. To some extent, government statistical departments have met the need of interview sample surveys by using the commercial market survey organisations already in existence in Sweden. For certain types of interview sample surveys, however, this has proved to be unsatisfactory, and special ad hoc organisations have had to be formed, involving considerable additional costs.

A working group<sup>2</sup> appointed on May 4th, 1950 by the Central Bureau of Statistics, (CBS) conducted an investigation regarding the need for a governmental organisation for interview sample surveys, and it subsequently proposed that a permanent organisation should be set up for this purpose. After certain additional investigations and preparatory work, for which a government grant was approved for the Fiscal years 1951-52 and 1952-53, the Swedish Parliament approved a sum of 100,000 kronor (5.18 kronor = \$1.), to be placed at the disposal of the CBS for the purpose of creating this organisation. The CBS also received funds to cover the salary of an administrative official, as the planned organisation was to be a division within the Bureau itself. In accordance with the resolution passed by the Swedish Parliament, the organisation, the official title of which is the Survey Research Centre, shall be self-supporting economically from the Fiscal Year 1954-55; the salary of the administrative chief will, however, continue to be paid directly from the CBS.

The Survey Research Centre will serve primarily as a service organ for government departments, committees, etc., for investigations and other work which require the service of a nation-wide net of representatives, and for consultation regarding statistics in general. An account of how the necessary machinery was set up for carrying out interview sample survey, etc., is given on the following pages.

---

<sup>1</sup> The opinions expressed here are not necessarily those of the Central Bureau of Statistics.

<sup>2</sup> This working group is consisted of Professor G. Enequist, Uppsala, Professor C. E. Quensel, Lund, Dr. E. von. Hofsten, Chief of Division, Social Welfare Board, Mr. I. Uhnöbom, Chief of Division, and Mr. F. Lublin, Chief Actuary, CBS, and the author, who was at that time an official at the Social Welfare Board.



## 2. CHOOSING THE FORM OF ORGANISATION

When planning a statistical investigation, the statistician is first faced with a choice between a total or a sample survey, and he selects the method which will give the required precision in results with the lowest possible costs.

If a sample survey is chosen, the same criterion is still applied for choosing between two different kinds of sampling systems:

(a) a system which (in its simplest form) can be described as one-stage sampling, using the elementary units of the population as sampling units.<sup>1</sup>

(b) a system which is characterised by the fact that groups of elementary units ("clusters") are used as sampling unit, ("cluster sampling").

One example of type of survey where system (a) is as a rule preferable, is when sampling is done from easily accessible registers which contain all necessary information. On the other hand, system (b) is usually more suitable for nation-wide interview sample surveys (and in the main any surveys where a large part of the costs refer to field work, and where these costs are greatly affected by the geographical spread of the field work).

A nation-wide interview sample survey is almost always based on two-stage sampling or on multi-stage sampling; (this may not be true for small countries like the Netherlands); usually, some kind of administrative units are used as first stage units (1-su's). By this means, a high degree of geographic concentration is obtained in the field work, as compared with what it would be when applying system (a).

The suitability of sample design also depends on the form of organisation for the investigation. The following two forms are considered here.

(A) The interviewers required for carrying out the field work are recruited locally, within each of the 1-su's. Each interviewer works only within the 1-su in which he resides.

(B) A corps of interviewers is recruited centrally and sent out to the 1-su's selected. (It is also possible to combine these two organisation forms by scheduling an interviewer to work in the 1-su in which he is resident, and in one or several 1-su's in the vicinity as well).

If the assignment is to estimate a certain total  $X$ , we express the variance function as  $\sigma_x^2$ , with  $x$  symbolising the estimator.

It is possible to write a relatively general cost function  $C(x)$ , which, with adequate numerical values of the parameters, can be used both for organisation forms (A) and (B). For instance,

$$C(x) = C_0 + C_1\sqrt{m} + C_2m + C_3mn + C_4m\sqrt{n}$$

is a relatively general cost function of this kind for two-stage sampling; here,  $C_0$  is an overhead cost,  $m$  the number of 1-su's in the sample design, and  $n$  the average number of second stage sampling units in each of the  $m$  1-su's.

If we decide to estimate  $X$  with a degree of precision given by  $\sigma_x^2 = \sigma_{ox}^2$ , we construct for each of the organisation forms (A) and (B) the function

$$F = C(x) + \lambda[\sigma_x^2 - \sigma_{ox}^2]$$

<sup>1</sup> The elementary unit is the smallest physical unit having the characteristic being studied

## THE CENTRAL BUREAU OF STATISTICS, SWEDEN

and determine the minimum value of  $F$ : we then choose the organisation form which has the lowest minimum (= lowest costs).

The components  $C_0, C_1 \sqrt{m}$  etc., which constitute  $C(x)$ , are usually of greatly varying relative significance for organisation forms (A) and (B) respectively. The overhead costs  $C_0$ , which inter alia include the costs for recruiting and training the interviewers, are thus in many cases considerably greater for (A) than for (B). In such a case, this can provide sufficient grounds for assuming that organisation form (B) is superior to (A) for a one-time investigation.

The situation is basically different if the overhead costs for organisation form (A) can be spread out over several surveys: in that case  $C_0$  is not so great for each individual survey and organisation form (A) is probably superior to (B). It is at this point that the idea of a general-purpose sample, (gps) of 1-su's comes into the picture.

The gps-idea means that one and the same sample of 1-su's is used for several surveys. By this means, it is possible, inter alia, to spread out the overhead costs over several surveys.

The gps-idea has proved to be especially suitable for an organisation for interview sample surveys which serves as a service organ for the whole of the governmental statistical production, as well as for commercial organisations dealing with market surveys.

The 1-su's which are included in a gps are selected with *known* probabilities. In principle, the sample can therefore be used for surveys dealing with any variables ( $X, Y, Z$  etc.). The effectiveness of the gps obviously is different for surveys of different variables and for the estimation of different parameters.

A gps which makes possible unbiased estimation at the time when the gps is constructed, retains this property for all time. It must, however, be taken into account that the effectiveness of a gps (as measured by variances), decreases with time, owing to the fact that the variations between 1-su's within strata (here it is tacitly assumed that the sample is selected from a stratified population) become successively greater. In this respect, the gps discussed here differs from the most common of all gps's—a table with random numbers.

In general, it can be said that a gps is suitable for statistical surveys of characteristics with high frequency and wide geographical spread; this applies both to surveys for the purpose of estimating parameters as totals and averages and, especially, to surveys for the purpose of estimating changes in such parameters. This situation is often found in surveys where the elementary units are individuals, families, households, dwellings, buildings, farms, etc.

For certain types of surveys, for which, according to the above, agps would be suitable, it must however be borne in mind that the gps alone may be insufficient, but it can be used to very good purpose. In general, this can be said to apply to surveys of variables with a marked skew distribution (retail trade sales are an example of a variable with marked skew distribution; a small number of large establishments account for a considerable amount of the total sales).

When, for instance, the total retail trade sales are to be estimated, a design of the following kind often gives a very good approximation to an optimum use of given resources; the population of business establishments is divided into three strata: large, medium-sized and small establishments (as regards sales). All large establishments are included in the sample, irrespective of their geographical position. The sample of medium sized and small



establishments is limited to the gps. Thus in each 1-su selected, all medium-sized establishments (= "locally large establishments" i.e. those which have not already been grouped with the very large), and a sample of the remaining small establishments are included. This technique is advantageous as soon as the survey aims at estimating the total of a variable with skew distribution; as a rule, the more skew the distribution is, the better the design.

Considering these points and studying the uses of gps in U.S.A., Canada and India, the working committee judged that organisation form A (i.e. using a gps) was the most satisfactory for a Swedish governmental organisation for interview sample surveys. It therefore recommended organisation form A for the CBS Survey Research Centre.

### 3. DESIGN OF THE GENERAL PURPOSE SAMPLE

Once the gps has been constructed, the statistician has limited possibilities of using the optimum principles given in the sampling theory; the choice is limited to different ways of sampling within the selected primary sampling units, and to the choice of the estimation method. Nevertheless, the sampling theory, in spite of the fact that it is a theory for choosing methods of estimating a *certain* parameter, can be made the basis for constructing a gps.

Let us consider the case of two stage sampling. The population includes  $M$  first stage units, 1-su's which we assume are of varying sizes. The  $i$ -th of the  $M$  1-su's thus includes  $N_i$  second stage sampling units; to simplify matters, we also assume that these  $N_i$  units are elementary units. The total  $X$  can be written as:

$$X = \sum_{i=1}^M \sum_{j=1}^{N_i} X_{ij} = \sum_{i=1}^M X$$

We select  $n$  elementary units by means of two stage sampling design, in such wise that the product of the sampling fraction  $f_1$  used in the first stage and that in the second stage  $f_2$ , is constant, i.e.

$$f_1 \cdot f_2 = f$$

The size of the sample will in this scheme be a random variable. As an estimator of  $X$ , we can now use

$$x = \frac{1}{f} \sum_{i=1}^m \sum_{j=1}^{n_i} x_{ij} = \frac{1}{f} \sum_{i,j}^n x_{ij}$$

where  $x_{ij}$  stands for sample observation in the  $i$ -th 1-su and  $j$ -th 2-su. This estimator is unbiased, i.e.  $E(x) = X$ .

The relative variance  $V_x^2 = \sigma_x^2/X^2$ , is given approximately by  $V_x^2 = V_{\bar{x}}^2 + V_n^2 + 2\rho_{\bar{x}n} V_{\bar{x}} V_n$

where

$$x = \frac{1}{n} \sum_{i,j}^n x_{ij}$$

and  $\rho_{\bar{x}n}$  is the correlation between  $\bar{x}$  and  $n$ .

As can be seen,  $V_x^2$  can be written as the sum of three components.

The first component  $V_{\bar{x}}^2$  is the relative variance for the estimator  $\bar{x}$  of the population mean  $\bar{X} = \frac{1}{N} X$ . This component in its turn consists of two part-components, one which refers to the variation *between* 1-su's and the other refers to the variation *within* 1-su's.

## THE CENTRAL BUREAU OF STATISTICS, SWEDEN

The second component  $V_n^2$  is the contribution to  $V_x^2$  from the variation in size  $N_i$  of the different l-su's over and above that already included in  $V_{\bar{x}}$ .

Finally, the third component  $2\rho_{xn}V_{\bar{x}}V_n$  is twice the product of  $V_{\bar{x}}^2$ ,  $V_n$  and the coefficient of the correlation coefficient  $\rho_{xn}$ . This  $\rho_{xn}$  is often very near to zero and then the following formula

$$V_x^2 = V_{\bar{x}}^2 + V_n^2$$

can be taken as a basis for continued discussion.

The formula for  $V_x^2$  shows that the variation in size of  $N_i$  of the l-su (for the sampling method chosen), makes a contribution to  $V_x^2$ ; this will be directly apparent in component  $V_n^2$  and indirectly in  $V_{\bar{x}}^2$ . (On the other hand, the variation in the  $N_i$  is of less significance when estimating averages (percentages)). Experience shows that  $V_n^2$  is often relatively great.

The problem of constructing a gps can now be formulated in the following way: the gps must be constructed so that the factors producing a high degree of variability in an estimator are given the least possible play. In a one-time survey, to be carried out by means of two-stage sampling, it is mainly in the following operations that this aim can be realised:

(a) construction of l-su's; it can be shown that the l-su's should be made as internally heterogeneous as possible, when only a small number of l-su's is to be selected.

(b) grouping the l-su's into strata: the strata should be made as internally homogeneous as possible; with this in mind, they should if possible be equal in size; in principle, the more strata which are constructed, the more effective is—*ceteris paribus*—the design.

(c) determining the probabilities  $P_{hi}$  for selecting the  $i$ -th l-su in the  $h$ -th stratum: The choice of the  $P_{hi}$  values, in proportion to some measure of the size of the l-su's, is an especially effective means of reducing the component  $V_n^2$ .

(d) selecting the l-su's: as a rule, the most effective design means selecting one l-su per stratum

(e) fixing the probabilities  $P_{hij}$  for selecting the  $j$ -th second stage sampling unit in the  $i$ -th l-su in the  $h$ -th stratum: here, too, it is possible to determine  $P_{hij}$  in an optimum fashion.

(f) method of estimation: the statistician usually has a choice of various estimators, some of which have less variance than others; as a rule, those which are least complicated from computation point of view are not those which have the least variance.

The construction of a gps is concluded by a number of l-su's being selected. As will shortly be explained, however, when constructing l-su's, stratifying them, etc., consideration must be given as to how the two operations (e) and (f) can be carried out when using the gps.

The plan used when constructing the CBS's gps is characterised by the fact that *quantitative numerical information (taken for instance from census figures), was to a certain extent reserved for choosing the estimation method, while qualitative information was to a great extent used for constructing l-su's and for stratification.* An account of how this plan was carried



out is given in the following, in connection with the account of how each of the above-mentioned operations (a)—(f) was carried out.

Further, when constructing a gps, and in contrast to what is the case when designing a sample for a single survey, it must be seen to that the design is as far as possible invariant in time. This aim can be realised if—in the first place when constructing the l-su's and when stratification is done—information is used which is subject to slight or no change during a period of, say, 10 years (geographical position is one example of this type of information); in addition, quantitative information which is most subject to change should be used for the estimation procedure.

*Some relevant information about Sweden:* Sweden covers a large area from North to South, lying between  $55^{\circ}20'$  and  $69^{\circ}4'$ , which corresponds to a distance of approximately 1,600 kilometres (1 km = 0.621 English mile). The total area is 450,000 square kilometres, of which 411,000 sq. kms. is land and 39,000 sq. kms. water.

Two of the many official divisions in Sweden, which are of special interest from the statistical point of view are the divisions into communes and parishes.

There are just over 2,500 parishes altogether, and one or more form a commune. The division into parishes forms the basis for the organisation of the population registration in Sweden.

There are 1,037 communes, of which 133 are cities or towns, 88 boroughs and 816 country district communes. Thus the country district communes, from the point of view of jurisdiction, total  $88+816 = 904$ .

These 904 country district communes are divided as follows, as regards area covered:

TABLE 1. AREA OF COUNTRY DISTRICT COMMUNES

sq. kms	no. of country district communes
- 99	218
100- 199	246
200- 499	308
500- 999	59
1,000-1999	42
Over 2000	31

On December 31, 1950, the total population was 7,074,039 persons (according to the 1950 census). Of this population, 3,296,241, or 46.8% lived in the real country district communes, 403,860, or 5.7% in the borough communes, and 3,343,938, or 47.5% in towns or cities. A total of 68.9% lived in built-up areas.

## THE CENTRAL BUREAU OF STATISTICS, SWEDEN

At that time, the 904 country district communes were divided as follows, according to inhabitants:

TABLE 2: POPULATION OF COUNTRY DISTRICT COMMUNES

no. of inhabitants	no. of country district communes
Under 1,000	3
1,000-1,999	68
2,000-2,999	247
3,000-3,999	248
4,000-4,999	133
5,000-5,999	82
Over 6,000	123

It is of interest to note that many of the country district communes which are large geographically have a small population.

On December 31st, 1950, the 133 town and city communes were divided as follows, as regards the number of inhabitants:

TABLE 3. TOWN AND CITY COMMUNES

no. of population	no. of town communes	respective share of entire town population
Under 2,000	4	0.2
2,000- 4,099	21	2.3
5,000- 9,999	37	8.2
10,000-19,999	36	15.3
20,000-99,999	32	35.4
100,0-and over	3	38.6
total	133	100.0

The total population of Sweden corresponds to a density of 17 inhabitants per square kilometre. Except for Iceland, Norway and Finland, this is the lowest figure in Europe. In Sweden, the density of population varies greatly for different parts of the country.

Over half, or 52.5% of the area of Sweden is comprised of communes having less than 5 inhabitants per sq. km. of land. Within this area, a total of only 405,561 persons reside—or 5.8% of the total population of the country. On the other hand 1,353,813 persons, or 19.2% of the total population reside in communes with at least 2,000 inhabitants per sq. km. of land; these communes however, only cover 0.1% of the total area of the country.

*Conditions to be met by the gps:* The CBS gps has been constructed to meet the following conditions:

It must be usable for nation-wide surveys, undertaken for the purpose of yielding figures covering the whole of Sweden.

It must also provide breakdowns for town and country districts.

The 1-su's must be so constructed that the population registers available can easily be made use of.

Each 1-su must constitute a geographically compact area.

The boundary between two 1-su's must in no case cut through a parish.

The four university towns must be included in the gps.<sup>1</sup>

The total number of 1-su's in the gps must be relatively small; the aim was set at "not more than 75". Out of the grant of 100,000 kronor made to the CBS for the setting-up of a Survey Research Centre, it was estimated that approximately 55,000 kronor would be used for recruiting and training a corps of interviewers. It was estimated that with two interviewers in most 1-su's and more in the larger 1-su's the funds would not permit more than 75, 1-su's.

The gps had to be such that it can be made the foundation of surveys based on probability sampling, further, it is necessary that this can be done simply.

*Type of information used:* The type of information used and their sources are given below:

(A) The 1945 population census: The information about population by age-group and by economic activity, and number of families with different number of children below 18 years for each parish, were used.

(B) 1950 population census. These figures have mainly been used for calculating the number of persons over 15 years in each parish.

(C) 1948 election results. The figures showing the division of votes between the Conservative, Liberal and Labour parties have been used.

(D) 1944 agricultural census. The number of farms in each of the following size groups:

0.26—10 hectares (small farms), 10-50 hectares (medium sized farms) and 50—w hectares (large farms) have been used.

(E) It should be noted here that while working out the gps, the CBS has been able to take advantage of consulting geographic labour market and other experts.

The construction of the CBS gps will be described below operation by operation.<sup>2</sup>

*Construction of 1-su's:* As mentioned before, the 1-su's should be made as internally heterogeneous as possible. In practice, the demand for heterogeneity means that the 1-su's

<sup>1</sup> By including the four university towns in the gps, opportunities will be improved for students of statistics, sociology etc., to get training in field interview methods, sampling methods etc.; arrangements for such training will no doubt prove helpful for the Survey Research Centre as well.

*The construction of the gps :*

<sup>2</sup> The sample design described here was worked out by the author in the winter of 1951, during a visit to the U.S. Bureau of the Census for the purposes of study, made possible by a grant from the Census for the purposes of study, made possible by a grant from the Economic Cooperation Administration. The author then had the advantage of being able to discuss a preliminary sample design, "A Design for a General Purpose Sample in Sweden", with sampling experts from the U.S..



are made large as regards the number of inhabitants included; this in turn leads to the 1-su's being large from the geographical point of view.

Most Swedish parishes can a priori be assumed as being too small for the purpose of effectively serving as 1-su's in a gps. The smallest administrative unit which could be considered was the commune; this idea was supported by estimates for certain variables in the population census of the contribution to the total variance from the variation between 1-su's consisting of parishes and 1-su's consisting of communes, in conjunction with cost data. Since communes do not always form markedly heterogeneous units, those in central and southern Sweden, where communes are considerably smaller than in the north of Sweden, and where communications are better developed, have been to a certain extent combined when constructing 1-su's.

In connection with this, where it was deemed possible from the communications point of view, agricultural and industrial communes, country and town communes, etc. have been combined. As for example, the first stage sample unit "Greater Uppsala" consists of the city of Uppsala, surrounded by a number of country district communes which are relatively similar to each other.

No reliable data on costs and variances have been available to make it possible to give an upper limit for the size of a 1-su with any satisfactory degree of precision. When in the future such data are available, it may prove that the 1-su's should have been made larger. In order to meet an eventuality of this kind, "super 1-su's" have been constructed.

Special points of view were borne in mind when constructing the 1-su's which included Stockholm, Gothenburg and Malmö. Thus one "Greater Stockholm", one "Greater Gothenburg" and one "Greater Malmö" were constructed (in the first two cases in cooperation with the two respective cities own statistical bureaus) to make it possible when making future surveys (possibly by extending their scope), to give separate data for these areas and further to make separate local surveys possible.

*Stratification:* The 1-su's thus constructed were then divided into 70 strata, homogeneous as far as possible. This stratification was done in the following manner.

Three 1-su's Greater Stockholm, Greater Gothenburg and Greater Malmö, each included more than 7,000,000:  $70 = 100,000$  inhabitants. They were therefore immediately classified as separate strata.

In 1945, these three separate strata together included approximately 1.3 million inhabitants. Thus the remaining 1-su's together included  $6.7 \text{ million} - 1.3 \text{ million} = 5.4$  million inhabitants, which were then to be divided into  $70 - 3 = 67$  strata; i.e. with approximately 80,000 inhabitants per stratum. Using a rule of thumb worked out by the U.S. Bureau of the Census in accordance with which 1-su's with approximately  $2/3$  of 80,000 inhabitants or more should each be made into separate strata (as with Greater Stockholm, etc.), 13 more 1-su's were classed as separate strata.

The remaining 1-su's constructed, were divided into two regions: one *expensive* region consisting of 1-su's in northern Sweden and western central Sweden; these 1-su's were characterised by the fact that the interviewing work within these 1-su's can be expected to be relatively *expensive*, because of the great distances involved and one *cheap* region, consisting of the remaining 1-su's: the interviewing work in this region can be expected to be less expensive than in the foregoing region.



The expensive region was then divided into  $L_E$  strata and the cheap into  $L_C$  strata  $L_E + L_C = 54$ .  $L_E$  and  $L_C$  were determined with the help of the formula:

$$L_E = \frac{F_E/\sqrt{C_E}}{(F_E/\sqrt{C_E}) + (F_C/\sqrt{C_C})} \cdot 54.$$

where

$F_E$  = 1945 population in the expensive stratum  
 $F_C$  = " " " " cheap " "  
 $C_E$  = cost per observation in the expensive stratum  
 $C_C$  = " " " " cheap " "

On the basis of available information, it was assumed that  $\sqrt{C_E}/\sqrt{C_C} = 2$ , when using the formula. The use of this formula involves an (attempted) approximation of optimum allocation in conjunction with optimum stratification.<sup>1</sup>

Statistical and expert information concerning the structure of economic life and general sensitivity of the market; geographical position, degree of urbanisation and size of the 1-su's intensity of industrialisation etc. were considered in stratifying the 1-su's.

The 1-su's which thus in 1945 had a high "agricultural percent" have in addition been characterised according to the structure of the agriculture. Thus in northern and west-central Sweden, a secondary classification was carried out on the basis of data regarding the division of the gainfully occupied population within agriculture and associated occupations into "essential agriculture", forestry and fishing. In the rest of Sweden, where "essential agriculture" dominates, the secondary classification has instead been based on the 1944 agricultural census data concerning the size of farming units.

The 1-su's which in 1945 had a high "industrial percent", were characterised on the basis of the 1945 population census data concerning the division of the gainfully occupied population within industry into the mining and metal industry, timber, paper and the graphic industry, food, textile and clothing industries, etc. Using these statistics, and with the help of information supplied by experts, a number of "industrial strata" were formed, the aim being that they should be characterised by homogeneity not only in comparison with the population census data, but also as regards market sensitivity and "risks of unemployment".

To a certain extent, the strata were constructed on the basis of data referring to branches other than agriculture and associated occupations and industry respectively. For instance, one stratum was constructed of 1-su's which can be characterised as large railway junctions and have therefore a high percentage of persons classed under the heading of "communications".

When carrying out stratification, it has been the aim to make geographically compact strata, i.e. the 1-su's which were placed in a certain stratum have all been selected within a limited area.

Where applicable, the aim has been when making the stratification, to construct strata which are homogeneous with regard to the proportion of town and country population.

<sup>1</sup> One possibility to deal with the cost problem would be to exclude a few very expensive 1-su's from the population actually sampled. Doing so would introduce some bias in the results of surveys; this does not necessarily mean that such an exclusion is not to be recommended: the savings may more than counteract the bias.

## THE CENTRAL BUREAU OF STATISTICS, SWEDEN

*Construction of "super-1-su's":* Since at present no data necessary for estimating the optimum 1-su size are available in Sweden, and it might therefore prove in future that the 1-su size chosen is too small (there is almost no likelihood that it is too large), the possibility has been allowed for, by means of constructing "super 1-su's", to extend the 1-su's if required.

A super 1-su consists of two (or in certain cases three) 1-su's adjacent to one another and belonging to the same stratum.

*Determining the probability of selection  $P_{hi}$  in each stratum:*

In the  $h$ -th stratum, the  $i$ -th 1-su was given a probability of  $P_{hi}$  to be selected. Obviously  $P_{hi}$  must be determined so that

$$\sum_i P_{hi} = 1$$

For the 1-su's which each form a separate strata,  $P_{hi} = P_{h1} = 1$ . In the other strata  $P_{hi} < 1$ .

For many surveys which may be carried out by the Research Centre, probabilities proportionate to the respective 1-su's population ("PPS") are satisfactorily close to the theoretical optimum  $P_{hi}$  value.

Amongst the measures for the number of persons included in the 1-su's, those over 15 years, in accordance with the 1950 population census, have been chosen. This number  $A_{hi}$  is approximately proportionate, for instance, to the total number of inhabitants, number of households (in different senses), number of dwelling units, etc. and probably to the total retail trade turn over within the 1-su in question.

*Selection of one 1-su per stratum:* In accordance with the plan drawn up for constructing the gps, one 1-su was selected in each stratum with a probability:

$$P_{hi} = \frac{A_{hi}}{\sum_i A_{hi}} = \frac{A_{hi}}{A_h}$$

The selection of a certain 1-su led automatically to the selection of the super-1-su, of which 1-su's were a part. This super-1-su was obviously selected with a probability proportional to the sum of  $A_{hi}$  for the 1-su's included in the super-1-su.

*Selection of "segments within" 1-su's:* The 1-su described as Greater Uppsala consists of a nucleus and a number of surrounding country district communes—"segments". The communes which form these, are all relatively similar; in other words, they can be regarded as a relatively homogeneous stratum. They have therefore been treated as a stratum on their own and one commune was selected. The reason for this procedure was, in brief, as follows. The communes in question are in many respects different from other agricultural communes in the county of Uppland; this dissimilarity is caused by the immediate vicinity of a big market. [Any of these communes can be reached from the town of Uppsala at little cost.] From the point of view of cost, it was not considered defensible for the present to include all these communes in a gps; one commune was therefore selected, but the possibility exists to select all of these communes. A similar method was applied in many other 1-su's which were constructed on the same basis.



*Use of the gps:* In principle, when making a sample survey which wholly or in part is based on the gps, the sampling fraction used in each l-su can be determined arbitrarily. In most cases, however, this is uneconomical and therefore unsuitable.

Two examples will be given here to show how the sampling fractions  $f_{2h}$  can suitably be determined when making different surveys.

For household surveys, it is often suitable to determine the sampling fractions  $f_{2h}$  within the strata so that:

$$f_{1h} \cdot f_{2h} = f$$

where

$$f_{1h} = A_h/A, f_{2h} = n_h/N_h$$

and  $f$  is a constant common to all strata, the over-all sampling fraction. If  $f_{2h}$  is determined in this way, the selection becomes self-weighting, which considerably facilitates the estimation procedure. Alternatively, one sampling fraction  $f_C$  might be used in strata where it is cheap to work and another  $f_E$  where it is expensive to work. In such a case, it is technically advantageous to make  $f_C$  an integer multiple of  $f_E$  (e.g.  $f_C = 2 \cdot f_E$ ); the total result is then obtained by simple weighing of the results from respective parts.

For a retail trade survey for the purpose of estimating the total sales, a suitable sample design can be formed along the following general lines.

All large establishments, irrespective of their geographical position in Sweden, are selected. For these establishments, the sampling fraction is then  $f_L = 1$ .

Of the medium-sized establishments, a sample is selected by including all such establishments situated in the l-su's which make up the gps. In the  $i$ -th l-su in the  $h$ -th stratum, the sampling fraction thus is  $f_{Mhi} = 1$  and for the  $h$ -th stratum as a whole, the sampling fraction for these establishments is  $f_{Mh} = P_h$ .

Finally, a sample of small establishments is selected by selecting a fraction  $f_{Sh}$  determined in such a fashion as to get a self-weighting sample of small establishments.

*Choice of estimation procedure:* This must be dependent on, inter alia, the sampling fractions used.

With the procedure recommended above for household surveys, i.e.  $f_{1h}$  and  $f_{2h}$  have been determined so that  $f_{1h} \cdot f_{2h} = f$ , where  $f$  is a constant common to all strata, a total  $X$  can be estimated by means of the estimator:

$$x = \frac{1}{f} \sum_h \sum_j x_{hj}$$

where  $\sum_h \sum_j x_{hj}$  is the sum of the  $x$ -variable for the units included in the survey.

In a retail trade survey for estimating the total sales  $X$ , in which the sampling plan described above is used,

$$x = X_L + x_M + x_S$$

is a simple estimator. Here

$X_L$  = total sales in the large establishments;

$x_M$  = an estimator of  $X_M$ , total sales in the medium-sized establishments, and

$x_S$  = an estimator of  $X_S$ , total sales in the small establishments.

# THE CENTRAL BUREAU OF STATISTICS, SWEDEN

As an estimator of  $X_M$

$$x_M = \sum_h \frac{1}{p_{hi}} \sum_j x_{Mhi}$$

can be used, and as an estimator of  $X_S$ :

$$x_S = \frac{1}{f_S} \sum_h \sum_j x_{Shij}$$

A few illustrations of the use of quantitative information available outside the survey, are given below:

Let us assume, that, by means of a labour force survey, we wish to estimate the number of persons employed. We can then estimate this number by means of the estimator  $x$  given above and estimate the total number of persons (= number employed + number unemployed),

by the estimator  $y = \sum \frac{Y_{hi}}{P_{hj}}$ ,  $Y_{hi}$  being auxiliary information available for the sample

1-su's. We now construct the estimator:

$$x_r = \frac{x}{y} Y$$

where  $Y$  is the actual total number of persons, calculated on the basis of population registers, for instance. The estimator  $x_r$  is an example of a ratio estimator; it is (as a rule) biased. This is — if certain relations between  $V_x$ ,  $Y_y$  and the correlation  $\rho_{xy}$  are fulfilled — superior to an unbiased estimator of the type

$$x = \frac{1}{f} \sum_h \sum_j x_{hij}$$

as it reduces the contribution to the total variance of the "variation between 1-su's". One example of a ratio estimator which (under similarly advantageous conditions) reduces the contribution to the variance of both "variation between 1-su's" and "variation within psu's" is the following estimate used in the U.S. Bureau of the Census' labour-force sample surveys, where  $x$  and  $y$  have the same significance as in the example above:

$$x'_r = \sum_a \frac{\sum_c \frac{x_{ac}}{f} \cdot \frac{Z_c}{Z'_c}}{\sum_c \frac{y_{ac}}{f} \cdot \frac{Z_c}{Z'_c}} \cdot Y_a$$

where

$x_{ac}$  = the sum of the  $x$ -variable in the sample for the  $a$ -th combined age and sex group in the  $c$ -th "colour-residence-group";

$y_{ac}$  = total number of persons in the sample in the same "ac-group";

$Z_c$  = total number of persons in accordance with the 1940 population census in the  $c$ -th "colour-residence-group";

$Z'_c$  = an estimate of  $Z_c$

$Y_a$  = total number of persons in the  $a$ -th age and sex group when making the survey;

$f$  =  $f_{1h} \cdot f_{2h}$ ; the sample is in other words self-weighing.



In a retail trade survey for the purpose of estimating the total sales  $X$ , suitable quantitative information could be used to construct analogous estimators. Data concerning the total sales for 1950 retail trade, in accordance with the 1951 Business Census, will probably prove very valuable in this respect.

The fact that certain quantitative information has been reserved for use in the estimation procedure, makes it possible, as results from new total censuses become available, (population, agriculture censuses, etc.) to use the results obtained and bring the gps up-to-date. To a certain extent, the data which formed the basis for constructing the 1-su's, stratification, etc., are not invariant in time; a 1-su which in 1945 was markedly agricultural, is perhaps no longer so in another ten years' time. This gradually leads to the fact that the contribution to the total variance from variation between 1-su's increases; if new total censuses are made, however, this tendency is counteracted by using the new information for the estimating procedure.

*Estimation of variances:* It is evident from the foregoing that the gps is the result of a selection of one 1-su in each of the 70 strata.

In principle, it is not possible to construct an unbiased estimator  $s_p^2$  of the variance  $\sigma_p^2$  in the estimator  $p$  of the parameter  $P$  (where  $P$  can be any parameter—for instance, a total or an average).

It is, however, possible to construct an estimator  $s_p'^2$ , which tends to overestimate  $\sigma_p^2$ . This possibility is based on the use of what the U.S. Bureau of the Census calls the "grouped stratum technique".

This technique is so named because the strata are grouped by using the principles for effective stratification, i.e., strata which are similar to each other are formed into groups which are as homogeneous as possible. The variance  $\sigma_p^2$  can then be estimated as in the case of stratified multi-stage samples with 2 (or more) psu's selected from each group.

The simplest way of applying this technique in the case above is to treat all 70 strata in this way, including those strata of one 1-su only. Another, more complicated method, is to deal separately with each stratum consisting of one 1-su and each consisting of more than one 1-su.

#### 4. SETTING-UP THE FIELD ORGANISATION

In 1954 a working group consisting of five persons including two temporarily attached sociologists was formed for the purpose of building up a field organisation. It was necessary to create a field organisation with about 150 local agents (this designation was used by the Survey Research Centre instead of "interviewers" to make clear the fact that their work involved more than interviewing only) and four district agents, in a very short time. This field organisation had been directed to become a permanent body for widely varying work in interview sample surveys, collection of prices for the index calculations carried out by the Social Welfare Board, listing, for instance, of business establishments, control of the completeness of aimed-at total surveys, and in tackling the non-response problem in mail surveys, for example.

*Recruitment of local agents:* In accordance with the plans for building up the field organisation, the working group was to recruit interviewers by contacting a number of persons directly, who could be considered suitable as local agents. For practical reasons,

## THE CENTRAL BUREAU OF STATISTICS. SWEDEN

contact had to be made by special persons on behalf of the working group, and the choice of these persons was in itself a considerable problem.

In recruiting *potential* local agents, attention was primarily to be paid to the applicant's personality (appearance, manners etc.) capacity (free time available for interview work) and estimated term of service as a local agent.

This system of recruiting by means of special contact agents, if rightly planned, gives a better result than recruiting by means of advertisement, etc. One condition for this is, however, that the contact agents are chosen, not only for being skilled themselves in making suitable contacts, but also because they are generally respected; they should, for instance, not be locally known as representatives of political parties or of fiscal or other authorities (e.g. the police authorities). As a basis for local recruiting, a written detailed list of conditions must be drawn up, giving the necessary qualifications for a potential interviewer. This list can be made into *guidance instructions* for the work, giving play for personal judgement to be exercised, rather than directions which must be followed strictly. The recruiting method chosen does not mean that it is unnecessary to make a selection among those who have been contacted and who are willing to become interviewers, but the final work of selection is less arduous than it would otherwise be. *Formal* education (examinations passed, etc.), should not be given undue emphasis. The ability to deal alertly with new problems is worth more than a wealth of book-learning.

A field organisation of the kind which the Survey Research Centre intended to form obviously had to include both men and women preferably so that male or female interviewers only could be used for certain surveys. Further, experience proved that interviewers must primarily give the impression of being *mature*; their age being a secondary matter.

In general, it can be said that these interviewers must be generally respected and be held in good esteem. A good rule-of-thumb is to avoid as interviewers any persons considered to be extremist in any way. Thus it is as a rule wise to avoid persons holding a supervisory office, as well as active representatives for any special cause.

The *capacity* of an interviewer can be judged with the help of information, including:

- (a) the applicant's working hours or (in the case of housewives, for instance) domestic duties;
- (b) place of residence within the interview district; as a rule it is an advantage if the interviewers live in the most populated part (if any particular district dominates) or centrally otherwise;

- (c) access to (public and private) means of communication;
- (d) the applicant's physique ("staying power"); since this as a rule is difficult to judge,

it is usual to set a maximum age limit for interviewers.

It is a matter of great importance to select persons for training as interviewers who can, as far as possible, be expected to give a long term of service. It is, however, less evident how it is possible to be able to give a good prognosis in this respect. This objective can be safeguarded to some extent by recruiting persons who have been given the opportunity of practical experience regarding the work involved as an interviewer. The risk of replacement appears to be particularly great amongst the younger unmarried persons. This is yet another reason for setting a relatively high minimum age limit (25 rather than 20 years).

The contact persons were requested to nominate at least four persons satisfying the following conditions.



(a) The agent's place of residence must be such that they can reach different places in the interview area easily and at low cost:

(b) When necessary, the agents must be able to devote a good deal of time to the work. It was specially emphasised that work might have to be done in the middle of the day and in the evenings, and not only on weekdays, but also at week-ends.

(c) It was pointed out further that in many cases, the work might require considerable physical and mental exertion. In view of this, the agents should not be more than 45 years of age, except in exceptional cases. Twenty five was given as a suitable minimum age. Only persons in good health could be considered; the Bureau stated that it did not consider persons with poor sight or hearing, or with any defect of speech suitable for agents.

(d) It was stated that persons with an elementary, grammar or similar school education were considered suitable. The same applies to persons who, by reason of office, union, or other work, are used to questionnaires.

(e) "Persons known in their home district as active representatives for the authorities, who in the eyes of the public hold a supervisory or similar position (e.g. police or tax authorities) or who actively represent certain associations, for instance, of a political nature, and who are in direct contact with the public in such a capacity", should as far as possible not be considered.

(f) Persons who in their work visit homes, business or similar establishments (e.g. commercial agents) should also not be considered.

In addition, each contact person was sent, for information purposes, a copy of an article, written by Dr. Karin Kock, Director-General of the CBS, "A Governmental organisation for interview sample surveys", published in the Statistical Journal 1952: 4.

Each proposed agent was required to fill in a form known as "personal application to the Survey Research Centre of the CBS". This was a simple questionnaire, to be filled in by hand about the personality, capacity, length of service etc.

The answers to this questionnaire made it possible to carry out the first test as to the suitability of the proposed local agent. Because of the fact that the questionnaire was to be filled in by hand, it was also possible to judge if the applicant's handwriting was legible—an important condition for a local agent. As far as was possible, the test was made uniform for all applicants, and led to the first selection being made; the fact that very few applicants were excluded witnessed to the fact that the work done by the contact persons was of a high quality.

The applicant left after the above selection had been made, then underwent a correspondence course, covering 5 letters. The course was divided as follows:

*Letter I.*

1. Origin of the Survey Research Centre, its work and organisation.
2. The need for and use of sample surveys.
3. Questions to supplement the form of application.
4. Test.

*Letter II.*

1. How a survey is made.
2. Different types of present-day surveys (interview sample surveys, collection of forms for postal surveys, listing of sampling units).

## THE CENTRAL BUREAU OF STATISTICS, SWEDEN

### *Letter III.*

1. Sample surveys and methods.
2. Questionnaire and interview technique.
3. Trial interview; example of questions.

### *Letter IV.*

1. Interview work.

### *Letter V.*

1. Interview work, continued.
2. General information.
3. Terms of payment.
4. Trial interview under realistic conditions.

In addition, each person taking part in the correspondence course received a memorandum concerning the work of collecting prices, drawn up by the Social Welfare Board; each person also had to collect details of a number of prices as a test.

The working group observed how each person taking the course dealt with the work; a person who (without an acceptable reason, e.g. illness) did not keep to the time plan of the course, was judged to be unsuitable as a local agent; since he (or she) would probably not keep to the timetable for certain work, if he were accepted as an agent.

The correspondence course included certain tests. Letter I was thus concluded with a clerical test, III with a simple trial interview, intended to be carried out with an acquaintance, V with a trial interview under realistic conditions.

The persons undergoing the correspondence course got a practical and theoretical insight into the work of a local agent. This proved to be particularly valuable; a large number of persons taking the course found on its completion that the work of a local agent did not suit them, and asked not to be considered for a post. By this "self-selection", the Bureau was saved much time and trouble the cost involved for such persons were relatively small compared with what they would have been if they had continued to the following personal course.

Those who had satisfactorily completed the correspondence course, were selected to take part in one of the four personal courses lasting six days, which preceded the final selection of the local agents.

*Training of district agents:* The four prospective district agents underwent a 12-day personal course at the CBS. The primary aim of this course was to train them for their future work, which can be described as follows:

(a) District agents must be contact agents between the Survey Research Centre and the local agents; furthermore, they must guide the local agents in their work and check it directly by subsequent contact with people supplying information, and by other means. In order to be able to carry out this work, the district agents must travel round to the local agents within their districts, and must also be accessible for them on the telephone at certain times.

(b) The district agents must also be the Centre's "first rank interviewers," carrying out more qualified work, for example, trial interviews, but must also carry out ordinary work and interviewing.



(c) When it is necessary, district agents must recruit and train new local agents themselves, and be able to replace temporarily any local agent, if so required.

(d) Finally, district agents shall be able to carry out other kinds of work, such as making selections of data from the population registers.

*Training courses for the local agents:* The aim of courses for local agents was to give final training to future local agents and to serve as the basis for the definite choice of local agents.

Special instructors were responsible for the work of the courses. These instructors included the staff of the Survey Research Centre, the four district agents, a sociologist, a statistician employed for a special survey and an official from the Social Welfare Board. In addition, an official from the Ministry for Social Affairs and one from the Social Welfare Board gave special lectures.

A far-reaching division of work was made the basis for the instructors' work, with only one person responsible for each subject dealt with (e.g. "interview technique", "administrative problems"), this guaranteed uniformity of instruction.

All courses were arranged with the participants living in the premises where instruction took place. This was done to reduce costs and to facilitate the work of the course and make it more effective. In addition, it was considered that this arrangement would help instructors and participants to know each other—an important aspect in creating a permanent field organisation. The course took the form of lectures and group and practical work.

Group meetings were devoted entirely to training for interview work; use was made of the role-playing technique, one acting as an interviewer and another as the person being interviewed, led by the instructor; moreover, the participants had opportunities to listen to recorded interviews (to demonstrate right and wrong interview methods).

Practical work included interview technique, collection of prices and listing. This work was preceded by special lectures. For practice in interview technique, a trial questionnaire was used, intended for use in a survey concerning 'old-age pensioners' living conditions, to be carried out by the Survey Research Centre. Each person taking the course had to carry out at least one interview (and as a rule carried out two), with an old-age pensioner selected in the commune in which the course took place. Experience thus gained was afterwards discussed at a joint and at group meetings.

Practice in collecting details of prices was acquired with about 10-15 commodities in 2-3 shops, and was carried out both in privately-owned and in cooperative shops.

Practice in listing was gained by making lists of shops for a sample of blocks.

Mealtimes and informal social gathering made it possible for participants and instructors to know each other more closely.

After the course was concluded, the instructors met and made the final choice of local agents. Those who were thus selected—just over 160 persons in all—signed an employment agreement.

*Continuation of training:* The CBS Survey Research Centre does not consider the training of local agents complete by their having taken the correspondence and personal course. In order that they shall maintain their standard of knowledge, they receive copies of articles on relevant subjects for due study. Thus they have received copies of the article on "Representativeness" published in the 1953 Statistical Journal, and on "The non-response

## THE CENTRAL BUREAU OF STATISTICS, SWEDEN

problem in statistical surveys". In addition, the agents by courtesy of the CBS and the Social Welfare Board, receive the papers "Statistical Journal" and "Social Information". Most important of all, however, is for the local agents to receive employment at regular intervals, as far as possible.

### 5. CONCLUDING REMARKS

It is natural to ask if the solution of the problem of building up an interview survey organisation as described above is good in the sense that the money invested will yield good dividends. It is, of course, too early by far to try to answer this question; instead we will only touch upon two relevant aspects.

Firstly: Is the sample of 1-su's chosen truly representative, and if so, will it continue to be representative?

As the 1-su's have been chosen with probabilities proportional to size, they are on an average larger than all the 1-su's from which they have been chosen. The important thing is that the sample of 1-su's will enable us to select samples of individuals, households or anything else which will give *representative results*. This characteristic of the sample of 1-su's is, in principle, time-invariant; it will not change in course of time. But it is true that the variances of estimates from surveys of a given size will probably be larger to-morrow than they are today. This increase in variance can be compensated for by increasing the size of the sample (of individuals, households or whatever are the secondary units); eventually however, we will reach a point where it does not pay to go on with the old set of 1-su's. Then it is time to consider the selection of a new sample.

Secondly: does it pay to invest 300 kr. in the recruitment and training of interviewers? The manner in which the field staff has carried through the first surveys (covering the period of activity from February 1—to May 1 1954) certainly justifies the statement that the training has proved effective. As an example of the performance, mention should be made of the unusually low non-response rate in the nation-wide survey on living conditions of old-age pensioners; the non-response was 0.8%. But this is not enough; the crucial point is how long will the interviewers stay with the organisation? It is too early by far to answer this question, but the prospects seem good: only 3 out of those originally recruited have left for such reasons as "getting tired" etc.; 3 others have left because they have moved out of the 1-su where they worked.

Thus, we will meet, sooner or later, the necessity of selecting a new sample of 1-su's; and we will face the necessity of recruiting and training new members of the field-staff. It seems natural to try to co-ordinate these tasks. This can be done as follows. When local agents resign, the work-load in the "empty" 1-su's will be taken over for a period by local agents from neighbouring 1-su's. Before recruiting new local agents, the sample will be partially revised. Such a partial revision is feasible as the present 70 strata can be grouped into a small number of geographic groups; the revision may therefore be limited to strata belonging to one such group.

### 6. ACKNOWLEDGEMENTS

Space will not permit listing the names of all the persons who have in some respect been helpful to me in my work on the construction of the sample described above, and building-up of the field-organisation. Such a list would include the names of a great many friends from India, England, the U.S., Canada and other countries, without whose support and stimulus I would never have been able to undertake the work.

## REFERENCES

- DALENIUS, T. (1951): *The Usefulness of an Organisation for Sample Surveys*, Stockholm, Mimeo., Swedish.
- (1953): Representativity. *Statistisk Tidskrift* 274-384.
- GOODMAN, R., and KISH, L. (1950): Controlled selection — a technique in probability sampling. *Amer. Stat. Ass.* 45, 350-497.
- HANSEN, M. H. (1947): The use of sampling in opinion surveys. *U.S. Bureau of the Census*. Washington.
- HANSEN, M. H. and HURWITZ, W. H. (1944): A new sample of the population. *Estadistica* 2, 483-491.
- (1946): The problem of non-response in sampling surveys. *Amer. Stat. Ass.* 41, 517-529.
- HANSEN, M. H., HURWITZ, W. N., and MADOW, W. G. (1953): *Sample survey methods and theory*. I-II, New York.
- KEYFITZ, N., and ROBINSON, H. L. (1948): The Canadian sample for labour force and other population data. *Population studies* 2, 427-443.
- REED, V. D., CAPT. KATHERINE G., and VITRIOL, H. A. (1948): Selection, training and supervision of field interviewers in marketing research. *The American Statistician* 2, 15-20.
- SHEATSLEY, P. B. (1950): An analysis of interviewer characteristics and their relationship to performance. 1-3. *International Journal of Opinion and Attitude Research* 4 (1950-51), 473-498; 5 (1951), 79-94, 191-220.

*Paper received : June, 1954.*



## RAGNAR FRISCH ON THE FACTOR COST CONCEPT OF NATIONAL INCOME

By HANNAN EZEKIEL  
*St. Xavier's College, Bombay*

In an article in *Sankhyā*<sup>1</sup> Ragnar Frisch has argued that the Factor Cost concept of National Income must be rejected because it implies an incomplete valuation of Government product. He asserts that those creative elements "that receive their remuneration through the special part of the government budget that has to do with indirect taxes and subsidies and similar items are being denied the status of 'factors'". This is incorrect. The full value of Government product is added separately to the private product under both factor cost and market price concepts. As between the two concepts, it is the question of the valuation of the private product which is really at issue. In the market price concept, this is taken at market value. In the factor cost concept, indirect taxes are deducted from the private product on the ground that these taxes never form a part of private incomes, having been simply collected on behalf of the State.

Frisch also insists that while national income should be invariant to purely institutional changes, "factor cost figures may be changed at will simply by shifting to another system of remuneration." This again is incorrect. The confusion arises because sufficient care is not taken to ensure that the changes are in fact purely institutional. A failure to obtain an invariant result in factor cost terms even when there are purely institutional changes is possible only if the full value of Government product is not added to the factor cost valuation of private product, as it should be. No doubt, the *distribution* of output between private and Government sectors will change on the imposition of indirect taxes, even when the total does not change. But this is as it should be.

Ragnar Frisch's attack on the Factor Cost concept is, therefore, invalid on both grounds. The Factor Cost concept neither fails to place the Government sector on the same footing as the private sector nor does it fail to satisfy the test of invariance.

---

<sup>1</sup> Frisch, Ragnar (1955): Market price versus factor cost in national income statistics *Sankhyā*. 15, 1-8.



## A REJOINDER

By RAGNAR FRISCH

*University Institute of Economics, Oslo*

If it were true that the full value of government product is actually included in the concept of national income at factor cost, then it would not be any difference between the national income at factor cost and the national income at market prices. The two figures would have to be identically the same. As a matter of fact they are not identical, which in itself is sufficient to show that Ezekiel's arguments cannot be correct in principle.

It is easy to explain why the full value of government product cannot be included: to do it, one would need to have in the first place an exhaustive theory of the way in which government activity is really functioning in a "productive" way and in the second place one would need an amount of statistical information which goes fantastically beyond anything which is at present available.

There would be only one way to include government product in a "full" way, namely to do it in the way I suggested in the numerical example appended to my *Sankhyā*-paper. But, of course, if one does it in this way, one will get back to a situation where there is no difference between the concepts of national income at factor cost and the concept of national income at market prices. Or rather the two concepts would be equal to the figures now designated under the name "national income at market prices".

## FOOD STATISTICS

By P. C. BANSIL

*The Socio-Economic Association, New Delhi*

Food statistics of India have come in for criticism so often. The findings of the National Sample Survey placing our food production at about 60 million tons in the year 1949-50 against the official estimate of 46 million tons excluding gram and 49.7 million tons including the same—was a big surprise to all impartial thinkers. A review of the progress during the first three years of the first Five Year Plan, if examined closely with regard to food, would throw a flood of light on the subject.

Food production in the country is shown to have increased by 12 million tons including grams and pulses. Let us see how much of this increase is due to development schemes, how much to providential factors and to what extent it is simply imaginary. Table 1 gives the necessary details about the increases.

TABLE 1. PRODUCTION OF FOODGRAINS

nature of foodgrains	production in million tons				
	1949-50 (base year)	1951-52	1952-53	1953-54	1955-56 (targets)
(1)	(2)	(3)	(4)	(5)	(6)
rice	23.2	20.7	22.5	27.1	27.2
wheat	6.3	5.8	7.4	7.8	8.3
other cereals	16.5	14.7	19.2	21.2	17.0
total	46.0	41.2	49.1	56.1	52.5
gram	3.7	3.2	4.2	4.6	—
other pulses	4.3	4.6	4.8	5.3	—
total foodgrains	54.0	49.0	58.1	66.0	61.6

It would be observed from the above that there has been an increase of 12 million tons in 3 years against the target of 7.6 million tons proposed to be achieved by various ways in 5 years *i.e.*, by 1955-56. This target was to be achieved by the various development schemes and each scheme was allocated a specific production target, *vide* Table 2 below.

TABLE 2. DISTRIBUTION OF PRODUCTION TARGETS  
BETWEEN VARIOUS SCHEMES

scheme	million tons					
	major irrigation	minor irrigation	land re- clamation	fertilisers and manures	improved seeds	total
(1)	(2)	(3)	(4)	(5)	(6)	(7)
additional production	2.01	2.38	1.51	1.15	0.56	7.6

It was assumed that about 0.5 million tons will be the contribution of community projects and National Extension Service scheme. The direct increase as a result of the various programmes was thus calculated to be only 7.1 million tons.

The Commission was fully conscious of the various seasonal factors. And it was clearly pointed out that allowances for seasonal variations cannot be made at this stage.

The year 1949-50, however, was a normal year for all intents and purposes. Any setback as a result of floods or draught or even the effects of an exceptionally good season were not to be considered in assessing achievements.

Before we go further into the matter, it would be necessary to find out the targets laid down for the various programmes and the achievements during the first three years of the plan period. Table 3 will illustrate this.

TABLE 3. PROGRAMME TARGETS AND THEIR ACHIEVEMENTS

programme	target for 1955-56 over the base	achieve- ments during 1951-54	percentage of (3) to (2)
(1)	(2)	(3)	(4)
major irrigation (million acres)	8.5	2.8	33
minor irrigation (million acres)	11.2	5.3	47
land reclamation (lakh tons)	14	8.16	58
fertilisers (thousand tons)	528.6	307	58
improved seeds	details not available		

In the case of development programmes, there is a time lag in the sense that the land reclaimed or water made available from an irrigation scheme in a particular year is put to productive use only the next year. In the case of fertilizers, the position is altogether different. There is nothing to guarantee that the amount supplied is used only for the particular crop.

During ten years of control, hardly any fertilizer was used for food crops. This was mostly due to the fact that their use was not economical for the peasant who had also a bias against them.

The peasant psychology is, however, changing and the use of fertilizers for food crops is increasing every day. Even in the case of other programmes, we have to find out how much benefit would go to food crops.

The crop pattern in India has remained practically static. If an average is taken for the last few years we find that food has occupied about 88% of the total cropped area and 90% of the total irrigated area. Table 4 brings it out clearly.

TABLE 4. SHARE OF FOODGRAINS IN CROPPED AND IRRIGATED AREA

(million acres)						
year	net sown area	area under foodgrains	percent- age of (3) over (2)	total irrigated area	irrigated area under food	percent- age of (6) over (5)
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1948-49	244.0	215.0	88	50.1	45.7	90
1949-50	283.2	248.6	88	53.7	49.1	91
1950-51	291.6	250.5	86	55.6	50.3	90



## FOOD STATISTICS

Working on the basis of percentages given in Table 4, we find that out of the total reclaimed area of 8.2 lakh acres, some 7.2 lakh acres would go to food. Even if the whole of this land has already been cultivated, it can yield a maximum of 2.5 lakh tons of foodgrains, calculating at the rate of 787 lbs. per acre, the average for standard yields of all the food-gains for the quinquennium ending 1946-47.

Similarly out of the total of 8.1 million acres which has been provided with irrigation facilities, nor more than 7.2 million acres can go to food production. According to the official yard-stick, irrigation facilities increase the yield at the rate of 1.5 ton per acre.

This can give an additional production of 1.5 million tons. Here also it is presumed that all the irrigation facilities have been taken advantage of, though this by no means is the case.

As for the fertilizer, manure and seed schemes, it is not possible to estimate correctly the results. According to the standard yard-stick one ton of fertilizer is supposed to give an extra yield of 2 tons of foodgrains.

The increase in the yield as a result of better seeds is taken to be one fourth that of fertilisers. This means that whereas one ton of fertiliser used in an area of about 15 acres would increase the yield by 2 tons, increased yields of the same area as a result of the use of improved seeds would be an additional production of 0.5 ton per acre.<sup>1</sup>

The combined effect of fertilizers, manures and improved seeds would, therefore, be to increase the yield by 3.5 tons per acre.

The actual area benefited by these schemes is difficult to estimate. The only important food crop in this connection is rice, which had some 4 lakh acres under the Japanese Method during the year 1953-54.

Even if it is presumed that an equal area under wheat has been supplied with similar facilities, the total increase in the yield on this account from 8 lakh acres cannot be more than 5 lakh tons.

Recapitulating the whole position we find that the real increase as a result of man-made efforts comes to 2.25 million tons: land reclamation 0.25 million tons, irrigation 1.5 million tons, and fertilisers as well as seeds 0.5 million tons. We may add another 0.75 million tons on account of community projects and better cultural methods. This would mean that out of an increase of 12 million tons, the Commission is responsible for only 3 million tons or 25 per cent.

The balance of 9 million tons is due to either seasonal factors, reclassifications, changes in the methods of estimation or increases in the reporting areas.

The year 1953-54, was no doubt an exceptionally good year. But such a big rise over the base year which was normal, cannot be accounted for seasonal factors. The real fact seems to be that our food yields have been underestimated in the past to the extent of 20 to 25 per cent.

The reason, in the early stages, was that land revenues and other taxes were based on the yield estimates. The peasant thus had the tendency of showing the figures as low as possible.

Then came the control period of about ten years, when both the surplus as well as the deficit States tried to underestimate their yields to save themselves from the rigours of procurement.

Now, for the first time these depressing factors have been removed. We have also made much improvement in the methods of estimating yields as a result of crop-cutting

---

<sup>1</sup> If the additional yield of 0.5 ton as a result of improved seeds is taken as spreading over 15 acres, the benefit per acre would work out to only 0.05 tons. But we have assumed that the increase is per acre so that any margin of underestimation in the official yard-sticks is eliminated. The National Sample Survey in its Crop Survey Wing is trying to assess the effect of various schemes on food production. So long as their findings are not finalised, we have to work on the basis of the existing data.



experiments. It would, therefore, be no wonder if as much as 50% of the estimated increase *i.e.*, about 6 million tons is merely due to these factors.

Nearly two-thirds (8.5 million tons) of the total increase is shared by rice and millets. Millets are mainly rain-fed and rice can also flourish well only under conditions of assured water supply.

South India which had suffered from the failure of the north-east monsoon for full five years had abundant rains during the last two years. An increase of 3 million tons or even little more because of these natural factors is thus not impossible.

The inescapable conclusion would, therefore, be that out of the total increase claimed hardly 50% represents real increase. And even out of this 6 million tons, as much as 3 million tons is due to the benevolence of nature.

In the face of all this, the Planning Commission concludes that "something like 5 or 6 million tons represent a more or less permanent gain, which will be retained in an 'average' year, but the rest is attributable to the good weather".

While discussing the extraordinary increase of 221% in the case of Rajasthan, the Commission made a reference to a possible inflation of figures due to better estimates, but seems to have totally ignored it in making the over-all estimate.

We have already seen that the permanent increase cannot under any circumstances be more than 3 million tons. It would be absurd to attribute an increase of 6 to 7 million tons to the benevolence of nature, especially when the base year itself had more than 3 million tons of increase because of this very factor.

This analysis shows that the additional yield of 6 million tons in the year 1953-54 represents as one big slice of underestimation in the base year which incidentally is the year over which the findings of the NSS extend. With a better and improved machinery for the estimation of our crop yields, no wonder if the total gap between the official figures and cannot be minced.

*Paper received: February, 1955.*

### CORRIGENDA

**Completeness, Similar Regions, and Unbiased Estimation—Part I :**  
By E. L. Lehmann and Henry Scheffé, *Sankhyā* 10, 305-340.

An error has been pointed out to us by Professor Walter L. Smith in Example 5.2 of our paper. The extension described on p. 326 of a function  $f(t_1, t_2)$  satisfying (5.6) and (5.7) for  $\theta = \theta_0$  does not in general satisfy (5.6) and (5.7) for all  $\theta$ . The extension by symmetry and periodicity should be applied to the function  $f(t_1, t_2) |t_2 - t_1|^{n-2}$ . This will satisfy (5.6). In order to satisfy also (5.7) we may restrict ourselves originally to functions  $f(t_1, t_2)$  which vanish outside a strip  $S_\epsilon : \epsilon \leq t_2 - t_1 \leq 1 - \epsilon$  for some  $\epsilon > 0$ . It follows from this that

$$[\epsilon/(1-\epsilon)]^{n-2} \leq E_\theta [f(T_1, T_2)]^2 / E_{\theta_0} [f(T_1, T_2)]^2 \leq [(1-\epsilon)/\epsilon]^{n-2}.$$

In the construction below (5.8),  $A_{\epsilon+}$  and  $A_{\epsilon-}$  should then be defined as the parts of  $A_+$  and  $A_-$  inside the strip  $S_\epsilon$ .

# INDIAN STATISTICAL INSTITUTE

## TWENTYFOURTH ANNUAL REPORT : April 1955 — March 1956

### PART 1 : CONSTITUTION AND ACTIVITIES OF THE INSTITUTE

1. The Indian Statistical Institute completed the twentyfourth year of its existence during the year under review (1 April 1955—31 March 1956). As we reach the quarter century mark of our institutional life we cannot help looking back at our modest beginnings. On 17 December 1931 the Institute was brought into being by a resolution of a public meeting held under the Chairmanship of the late Sir R. N. Mookerjee who became our first President (1932-1936). It was duly registered in April 1932 as a 'non-profit learned society' under Act XXI of 1860. The First Annual Report of the Institute revealed a paid staff of a single part-time worker and a total current expenditure of Rs. 238 only. Since then its activities have become more and more diverse and extensive.

2. Broadly speaking three successive stages may be distinguished in its development. During the first few years of its existence the Institute functioned more or less as a laboratory for analytic studies including the design of experiments, and also took up small scale enquiries on behalf of Government departments and private concerns. During the second stage the Institute carried out, on an increasing scale, crop estimation surveys on behalf of the Governments of Bengal and Bihar and developed the technique of large scale sample surveys for this purpose. Since 1950 the Institute has taken a leading part in organizing the National Sample Survey at the desire of the Government of India, and since 1954 is actively helping in the work of planning for national development. Two new units were established during the year under review, one called *Kalyanashree*, a production centre at head-quarters in Calcutta for the study of the economics of household handicrafts and industries under experimental conditions, and the other, the Industrial Management Research Unit for Planning (IMRUP) at Bangalore.

3. *Constitution and Administration*: The Institute consists of Ordinary, Life and Honorary Members; Associate Fellows, Fellows and Honorary Fellows. The supreme control including the power of making rules is vested in the members in General Meeting assembled. The President and Office-bearers are elected annually. The management is vested in the Council, the Governing Body of the Research and Training School, and other Committees elected from time to time. Work assigned by the Government of India is done in accordance with conditions settled by mutual agreement and in consultation with Government.

#### ACTIVITIES

4. The activities of the Institute cover a wide range and may be grouped under the following heads:

1) *Research and Training School (RTS)*: For research in theoretical and applied statistics with sections for a 3-year course of professional training at postgraduate level for candidates who have already taken their master's degree and other technical courses at various levels. It has three units for biometric, anthropometric and psychometric researches



attached to it. During 1955-56 it received a grant of Rs. 7,26,000 from the Government of India for this purpose.

2) *Examinations*: Since 1938 the Institute has been awarding Statistician's Diplomas and Computer's Certificates. It has also arrangements for the award of Associateship and Associate Fellowship of the Institute on the basis of professional qualifications.

3) *Projects*: For statistical enquiries and investigations which are undertaken on the basis of ad-hoc grants. The biggest project the Institute has been handling since 1950 is the technical work of the National Sample Survey.

4) *Electronic Computers*: These activities cover both development and construction work in connection with electronic (analogue and digital) computers, desk and other types of calculating machines, precision measuring instruments and associated equipment. The Institute has an electronic laboratory and a well-equipped workshop. The Institute has, of late, been receiving electronic and other equipments from USSR through the United Nations.

5) *Statistical Quality Control (SQC)*: Three whole-time SQC Units are maintained at Bangalore, Bombay and Calcutta which work under the guidance of a Policy Advisory Committee.

6) *Operational Research relating to Planning*: An Operational Research Unit (ORU) started work on a small scale at the end of 1953. In November 1954, Prime Minister Nehru inaugurated at the Institute studies relating to planning which led to the preparation of the Draft Plan-frame of March 1955. The Institute in association with the Central Statistical Organization, the Ministry of Finance and the Planning Commission, is continuing the work on perspective planning.

7) *General Services*: The Institute maintains common services in the form of a large Machine Tabulation Section and a specialized library with a Photographic Section. There is also an Estate Office to look after construction, repair and maintenance of buildings and water, drainage and electricity.

8) *Social Services*: The Institute offers various social services and amenities to its workers, students and guests such as medical welfare, hostels, guest houses for visitors, canteen, night school, transport service and workers' club for sports, recreation and cultural activities. An adult literacy drive has recently been launched among the workers of the Institute.

5. The following associated bodies were organized on the initiative of the Indian Statistical Institute and work in close relation with the Institute:

1) *International Statistical Education Centre*: This is an associated institution established in 1950 under the sponsorship of UNESCO and is being maintained under the joint control of the International Statistical Institute and the Indian Statistical Institute with the support of the Government of India.

2) *Statistical Publishing Society*: This Society was established in 1935 on the initiative of the Institute to undertake the publication of *Sankhyā*, the Indian Journal of Statistics as the official organ of the Institute. The Society has under its control the Eka Press which is well equipped for undertaking scientific and technical work of a high quality and is managed in close association with the Institute.

## TWENTYFOURTH ANNUAL REPORT : 1955-56

3) *India Calculating Machine and Scientific Instrument Research Society*: This is a 'non-profit' society established on the initiative of the Institute and registered in 1943 under Act XXI of 1860 with the object of promoting research, study, production and use in India of calculating machines, statistical, mathematical, scientific and engineering instruments, apparatus and appliances of all kinds.

6. The Institute, of course, still functions as a learned society with *Sankhyā*: The Indian Journal of Statistics, as its official organ and has society type branches at Aligarh, Bangalore, Bombay, Madras and Poona. Besides, it has today many other activities for which operating offices are maintained at Baranagar (Calcutta 35), at three places in the city of Calcutta, and at Bombay, Bangalore, Delhi and Giridih (Bihar). The staff increased to over 1400 workers in December 1955 and the annual expenditure went up to fiftytwo lakhs of rupees in 1955-56.

7. The main offices of the Institute are at present located on the Institute's own land at 203 Barrackpore Trunk Road in a seven-storeyed building covering 62,400 sq.ft. of floor space. In addition about 30,000 sq. ft. of floor space is occupied by the Institute at 202 Barrackpore Trunk Road, acquired by the Government of India for the Institute's work. The Institute also uses about 50,000 sq.ft. of space in other premises in Baranagar and Calcutta. In Giridih, the Institute acquired about 35 acres of land for location of an experimental farm.

### STATUS AND CONSTITUTION

8. The Indian Statistical Institute had started receiving grants from the Government of India in 1936; and practically since then the question of the stabilization of the Institute has been under consideration by the Central Government. Many different approaches have been explored for this purpose during the last twenty years. From time to time there were proposals from Government to take over the Institute, or to convert the Institute into a primarily educational and research institution. After several years of negotiations new rules were framed in 1950, at the desire of the Ministry of Education, in which the teaching activities were made the main function of the Institute under the control of a Governing Body with Government representatives. The new rules could not, however, be brought into effect owing to certain technical difficulties; and in 1951 the provision for the Governing Body was incorporated in the constitution in the form of regulations and the Council retained its status as the general co-ordinating agency.

9. In 1952 the Cabinet of the Government of India decided that the Institute should function as a focal centre for advanced studies and research in statistics in India; and since then extensive programmes of professional training have been developed jointly by the Institute and the Central Statistical Organization. In 1953 there was a proposal that the Institute should be converted into a University under a Central Act, but it was felt that there would be difficulties in preserving the operational and society type activities within the framework of a University. Since 1954 the Institute has been increasingly participating in studies relating to planning for national development. It was agreed that the Institute should retain its autonomous status but would be recognised by the Government of India as an institution of national importance. The rules of the Institute were substantially changed in April 1955.



## APPENDIX : HISTORICAL BACKGROUND

1. The beginnings of the Indian Statistical Institute were modest enough : a solitary computer working part-time and a total current expenditure of rather less than Rs. 250— that was all when it started in April, 1932. In September 1956, the Institute had on its roll 1674 regular workers who were backed by an annual budget allocation of Rs. 52,00,000. Impressive as these figures are, they convey little idea of the nature of the work that is being carried on by the Institute, its wide range, its complex character and the specialization at high scientific levels that is necessary for carrying out investigations that have an intimate bearing on the life of the nation.

2. A long period of preparatory activity, going back to the early years of the first world war, preceded the formal inauguration of the Institute. Prasanta Chandra Mahalanobis was back in India in 1915 on a short holiday after he had been awarded a senior scholarship of King's College, Cambridge, for research in Physics. Just before his departure from Cambridge, his tutor, W. H. Macaulay, had drawn his attention to the *Biometrika* and Karl Pearson's *Biometric Tables*, copies of which he brought with him to India and researches. But he did not return to Cambridge. Accepting the post of a professor of Physics in the Presidency College, Calcutta, he decided to stay back in India and carry on his statistical studies alongside of his work as a teacher of Physics. In Calcutta, he came into intimate contact with the eminent scholar and philosopher, Brajendra Nath Seal, who had a full appreciation of the importance of modern statistical methods and encouraged Mahalanobis to take up this subject as his life's work.

3. Urged by the great savant, young Mahalanobis entered upon his statistical career with gusto. Not content to confine his interest in statistics to merely academic studies, he was soon applying statistical methods for the solution of problems in anthropometry, which led to the formulation of the Generalized Distance ( $D^2$ -statistics). In meteorology, he found ample scope for trying out his new-found tool, enabling him to locate, at a height of about four kilometers above sea level, the region of highest control for changes in meteorological conditions on the surface of the earth; later, this result was confirmed by Franz Baur from physical considerations.

4. More far-reaching from the point of view of national welfare were the results of extensive investigations into rainfall and floods in North Bengal, Orissa, and West Bengal. These led not only to suggestions for really effective flood control but also supplied in later years some of the basic calculations for two great river valley schemes, namely, the Damodar Valley and the Hirakud Projects.

5. Meanwhile, Mahalanobis had gathered around him several part-time assistants and a group of young research workers to form the nucleus of the Statistical Laboratory which was located in the Presidency College, Calcutta. Amongst this pioneer band were Sudhir Kumar Banerjee, Subhendu Sekhar Bose, Nistaran Chakraburty and Jaladhar Sarma, the only one amongst the four who is still serving the Institute. Nistaran Chakraburty is now the Director of the West Bengal Statistical Bureau. Sudhir Kumar who became the Chief Computer and Subhendu Sekhar, who was the leading spirit amongst the workers of the Institute, have both been cut off by untimely death in the prime of their careers. In them, the Institute has lost two of its most ardent pioneers.

6. The Laboratory also had been gradually developing. By 1930 it owned about Rs. 25,000 worth of books, calculating machines and other equipment; and had undertaken some enquiries on behalf of the Government departments and private concerns. The first official recognition came in July 1931 with an annual grant of Rs. 2,500 from the Imperial (now Indian) Council of Agricultural Research for statistical investigations relating to agriculture.

7. The year 1931 marks another milestone in the progress of statistical organization in India. At a meeting convened by P. N. Banerjee (at that time the Minto Professor of Economics), N. R. Sen (Ghosh Professor of Applied Mathematics) and P. C. Mahalanobis (Professor of Physics, Presidency College) and held on 17 December 1931, with Sir Rajendra Nath Mookerjee in the Chair, a resolution was unanimously adopted to establish the Indian Statistical Institute. The constitution was approved at the end of February and the Institute was formally registered as a non-profit learned society in April 1932. Society type branches were established quite early in Mysore, Poona, Bombay, Madras, Lahore, Banaras, Lucknow and Delhi; and the Statistical Laboratory in Calcutta continued to function as the active nucleus of the Institute.

8. This was how the Indian Statistical Institute was formally organized. The first general grant for research and training came in 1935 in the shape of Rs. 5,000 a year from the Government of India. A big development occurred when the Institute took up in 1937 a five-year project, on an expanding scale, to develop a sampling method for estimating the acreage and out-turn of the jute crop in Bengal with the required accuracy at a reasonable cost. Since then, the Institute has been a foremost centre of sample surveys in the whole world. From this time, the larger part of the income of the Institute began to be derived as contract grants for applied projects and enquiries.

9. The training of officers deputed by Central and State Governments had started from 1932. In the earlier years such training was more or less on an individual basis; and over 150 individuals received such training at the Institute between 1932 and 1939. From 1939, the Indian Statistical Institute started organized courses of instructions and the training courses gradually developed into the present Research and Training School at a post-graduate level.

10. A scheme for examinations for the award of certificates for computers and diploma for statisticians was got ready in 1935. These examinations, which were started in April, 1938, served a real need, and have become quite popular and are held all over India every year.

11. To offer facilities for the publication of research papers in India, it was decided in 1933 to start *Sankhyā*: The Indian Journal of Statistics, as the official organ. On the initiative of the Council of the Institute, arrangements were also made to establish an associated non-profit institution, the Statistical Publishing Society to maintain the Eka Press and publish *Sankhyā*, which soon became a journal of international recognition.

12. The Indian Statistical Institute had been for some time pleading for a separate statistical section to the Indian Science Congress. Such efforts, however, did not bear any fruit and so in 1938 the Institute organized on its own the First Session of The Indian Statistical Conference which continued for several years. The Indian Science Congress agreed to start an independent section for Statistics from 1946.



13. Ventures in research and project work along with promotional activities soon earned international recognition for the Institute and helped to enlisting active association with it of distinguished foreign statisticians who came out to India to work and lecture in the Institute as visiting professors. The first to come was Professor (now Sir) Ronald A. Fisher, who paid his maiden visit in 1938, repeating it in 1945, 1951 and 1954. Other visits have followed and except for an unavoidable break during the Second World War, a stream of distinguished foreign scientists, statisticians as well as specialists in other fields of science, have responded to the invitation of the Institute to come here and work as research workers and guest lecturers.

14. The Institute has thus become a real international centre of research as well as a forum for scholars and scientists from all over the world to work and hold discussions. Mention may be made of a few amongst these distinguished visitors, such as, R. A. Fisher, J. B. S. Haldane, J. R. N. Stone, F. Yates of the UK; J. K. Galbraith, Harold Hotelling, W. Hurwitz, Simon Kuznets, Walter A. Shewhart, Abraham Wald, Norbert Wiener of the USA; D. D. Degtyar, V. A. Ditkin, Y. V. Linnik, I. Y. Pisarev, M. I. Rubinstein of the USSR; C. Bettelheim (France); Tosio Kitagawa and Motosaburo Masuyama (Japan); J. Tinbergen (Netherlands); Ragnar Frisch (Norway); Oskar Lange (Poland); H. Wold (Sweden) and Arthur Linder (Switzerland).

15. After some initial dislocations caused by the Second World War, the project side of the Institute developed rapidly, as a result of the increasing need of statistics by Government. The Sample Survey in Bengal was extended in 1943 to cover both the area sown and the total yield of jute, rice and other important agricultural crops throughout the year. The tabulation on the basis of the two per cent Y-sample of the 1941 Population Census started in 1944-45, and various other socio-economic surveys were undertaken.

16. The post-war years have witnessed a rapid expansion and many new developments. The question of stabilization of the Institute had been under consideration of the Government of India since 1938. Shri Chintaman Deshmukh, who has been the President of the Institute since 1945, helped in solving many difficulties; and it was mainly through his efforts that the Research and Training School was established with an initial recurring grant of Rs. 4.5 lakhs from the Government of India from 1949-50. Administrative sponsorship of the Institute was transferred from the Ministry of Education to the Ministry of Finance in 1950, and to the Cabinet Secretariat of the Government of India in 1956. A Governing Body was established in 1952 to look after the affairs of the Research and Training School, subject to general co-ordination of the work of the Institute as a whole by the Council.

17. From about this time, the Indian Statistical Institute began to emerge as a national institution. From 1950, the Institute started working on a vast project, namely, the design and analysis of the data of the National Sampling Survey which is collecting comprehensive information relating to social, economic and demographic characteristics on a countrywide basis in the form of two "rounds" of survey every year covering both rural and urban areas. This is reputed to be the biggest sample survey of its kind in the world today.

18. In 1950, the Institute also helped to bring into being, in collaboration with the International Statistical Institute and under the sponsorship of the UNESCO, the International Statistical Education Centre, which is being maintained with the financial support of the Government of India. Since then, the ISEC in Calcutta has been providing statistical training to students from many Asian countries; about 202 trainees have come up to 1956.

## TWENTYFOURTH ANNUAL REPORT: 1955-56

19. The work of the Institute began to receive increasing recognition both in India and abroad. Professor Mahalanobis was elected a Fellow of the Royal Society in 1945, and acted as the Chairman of the United Nations Sub-Commission on Statistical Sampling from 1947 to 1951; he has been a member of the UN Statistical Commission from 1946 and Chairman from 1954. Other workers of the Institute enjoy international reputation and some of the old members of the staff are working in other places in and outside India. The Institute acted as the host society to the International Statistical Conferences in India in 1951.

20. Since Professor Mahalanobis started working as Honorary Statistical Adviser to the Central Cabinet in 1949, the Institute is becoming more and more closely associated with the Central Statistical Organization (CSO) and other Government agencies in New Delhi. From 1952, the Institute is functioning as the focal centre for professional training and research and as a National Statistical and Computational Laboratory.

21. An associated non-profit institution for the development of calculating machines and scientific instruments had been established in 1943 with the approval of the Council of the Institute and a small workshop was started which, however, did not make much progress in the beginning. A small workshop was started in the Institute in 1950 for the repair and maintenance of desk calculating machines and other equipment a little later: and is now in operation. An electronic laboratory was started a little later where a small electronic analogue computer was designed and constructed in 1953. An electronic digital computer was purchased. In 1955, arrangements were made to secure a bigger digital computer and a large number of machine tools from the USSR, through the United Nations Technical Assistance Administration and further developments are in progress.

22. The Institute has been promoting the introduction of Statistical Quality Control for a long time; and had started the earliest courses in this subject in India in 1945-46. A separate SQC Section was established in 1953 which has whole-time units at present working at Bangalore, Bombay and Calcutta.

23. Still more significant were the developments in 1954. The Indian Statistical Institute was called upon by the Planning Commission to undertake, jointly with the CSO, work on national planning with a dual objective of solving the unemployment problem in 10 years and continuously increasing the national income as rapidly as possible. In November 1954 Prime Minister Nehru inaugurated in the Institute studies relating to planning for national development. This work was carried on with the active collaboration of the CSO, the Department of Economic Affairs (Ministry of Finance), and the Economic Division of the Planning Commission. Many of the foreign experts who came to the Institute also participated in the studies and discussions on planning.

24. On the basis of the above work, the "Draft Plan-Frame" was submitted to the Prime Minister in March 1955 and was accepted in May, 1955 by the National Development Council as the basis for the formulation of the Second Five Year Plan. The Institute thus became intimately connected with national planning in India.

25. The activities of the Institute are diverse and developments have occurred and are still occurring in many directions but there is a unity in its purpose. The Institute has been trying for twentyfive years to promote national development through the patient collection and analysis of statistical and technical information and their utilization for policy and administrative decisions in a scientific manner.



## PART 2. YEAR UNDER REVIEW : 1955-1956

## 1. RESEARCH AND TRAINING SCHOOL

1. The main activities of the Research and Training School fall under the three heads—(i) Research, (ii) Consultation and (iii) Training.

## RESEARCH

2. Research has been conducted at various levels: (a) theoretical research on the foundations of probability and the logic of statistical inference, (b) applied research involving the development of statistical tools for application to specific problems, and (c) research in fields of application in which the approach is statistical.

3. *Research Seminars*: Besides the seminars regularly conducted by the Research and Training School for exchange of ideas among the staff, special courses of lectures on recent advances in mathematics and statistics were arranged. The first was a series of 60 lectures of Harmonic Analysis and Multiple Prediction Theory by Prof. Norbert Wiener of the Massachusetts Institute of Technology, USA. Twenty participants attended Prof. Wiener's seminars in response to the Institute's offer of financial assistance for this purpose to interested research workers all over India.

4. Continuing the seminar lectures started in 1954-1955 on some problems related to planning in India, Prof. P. C. Mahalanobis developed the basic statistical and computational problems involved in the formulation of a plan for economic development. He gave a four sector model for determining the investments in (1) producer goods, (2) factory production of consumer goods, (3) cottage industries and (4) services (health, education etc.) and explained its usefulness in solving the problems of planning in India to achieve the main targets: (a) liquidation of unemployment in 10 years, and (b) doubling of national income in 14 years.

5. *Theoretical Research*: The research work carried out in theoretical statistics by the staff and research scholars of the School has been on the following topics:—Design of Experiments, Tests of Hypotheses, Estimation Theory, Sample Surveys, Stochastic Processes, Distribution Problems, Mathematical Models for Economic Planning, Industrial Sampling, High-speed Computation, and Problems of Optimum Selection. About 35 research papers on these subjects have been prepared during the year. Some have already been published or are in the press, while others have been preliminary reports. A summary of research work done by the School is given in Appendix 9.

6. *Applied Research*: Work in applied research by the School was mainly undertaken by the Biometric, Anthropometric and Psychometric Units and members of the staff attending to scientific enquiries.

(i) *Biometric Unit*: This unit for applied research in biometry was started in April 1954 under the leadership of Dr. M. Masuyama who returned to Japan in August 1954. Professor J. B. S. Haldane, F.R.S. and Mrs. Haldane (Miss Helen Spurway) arrived in July 1954 and worked in the Biometric Unit for about two and a half months. The unit is now under the charge of Dr. B. C. Das who joined the Institute in October 1955. Besides Dr. Das, the Unit has six professional workers including a biochemist and a medical practitioner. A well-equipped laboratory has now been set up which possesses among other things,

## TWENTYFOURTH ANNUAL REPORT : 1955-56

electrophoresis with scanning and other accessories, direct reading meter, magnum centrifuge, incubator, thermostatic bath, distilled water plant, air conditioner, microscopes etc. Some of the activities of the unit have been study of worm-activity on soil, experiments on fish culture, clinical and blood group tests, etc.

(ii) *Anthropometric Unit* : A good deal of anthropometric studies were carried out in the Institute in the early days by Professor Mahalanobis and later by Dr. C. R. Rao but there was no provision for systematic work. During the year, an Anthropometric Unit was started with two anthropologists who had been trained by Dr. D. N. Majumdar, Professor of Anthropology, Lucknow University and had been doing research for some considerable time under his guidance with the help of Institute research fellowships. This Unit is currently engaged in techniques of data collection such as of measurement, standardization of instruments, etc.

(iii) *Psychometric Unit* : This has been an year of rapid growth for the Psychometric Research and Service Unit established in 1954. The staff of the unit has increased from three to ten during this period. The unit's major activity was concerned with selection tests. A number of tests have been constructed, administered, scored, reported, and interpreted in connection with the selection of job-applicants, trainees, college students etc. Besides, some theoretical problems which arise in testing projects, like correction for 'guessing', validity coefficients for 'restriction of range' and its extension to 'two-stage' selection, etc., have been investigated. The Unit continues to edit the Psychometrics section of the 'Psychology News Bulletin', which reports work in progress or recently completed in India.

### CONSULTATION

7. *Scientific enquiries* : Consultation has been open to all scientific workers in the country. Advice has been given to various enquiries on the planning of investigations and analysis and interpretation of data. The Research and Training School attended to a number of scientific enquiries from research workers in various fields, government departments and business firms. Some of the enquiries handled are listed in Appendix 5 and included such problems as (i) correlation between mental age and scores in arithmetic of secondary school students; (ii) analysis of manurial and varietal experiments on paddy; (iii) statistical analysis of the effect of a new drug on cholera; (iv) estimation of linkage between certain factors in paddy.

### TRAINING

8. During the year under review there has been increased demand for training in statistics and considerable expansion in the training programmes of the School, both in the number of courses conducted and the number of trainees attending each course. The two-year advanced training course has been converted with effect from July 1955 into a three-year advanced theoretical-cum-professional course to meet the rapidly increasing demand all over the country for statisticians adequately trained in professional work in a large variety of jobs. A special training course in statistics of six or nine months' duration for officers deputed by the Central and State governments and recognised institutions was organized jointly with the Central Statistical Organization (CSO), New Delhi. A short-term evening course for statisticians has also been introduced to suit the convenience of persons who are already in employment and who intend to acquire knowledge of basic statistical methods for use in their actual work.



9. The different courses of training that have been given during the year including the new courses mentioned above have been :

- (a) Three-year Statistician's Course,
- (b) Short-term Statistician's Course,
- (c) Officer's training (jointly with the CSO),
- (d) Officers on deputation,
- (e) Training for Computers.

In addition, training is also given at the International Statistical Education Centre.

10. *Advanced Studies* : Research scholarships were awarded to 12 students for carrying out work on subjects such as Advanced Probability, Stochastic Processes, Statistical Inference, Multivariate Analysis, Statistical Quality Control, Psychometry, Biometric Methods etc. (List in Appendix 11). During the year under review, three workers were admitted to the degree of Doctor of Philosophy of the Calcutta University and two others submitted theses for the same.

11. *Training courses* : (a) *Three-year statistician's course* : As already mentioned, the two-year statistician's course has been converted into a three-year course with effect from this year. This training is arranged in collaboration with the CSO, where the trainees spend from three to six months for training in official statistics. A special feature of the new course is that only highly qualified students i.e., those with at least a good Master's Degree are admitted ; no tuition fee is charged and the trainees receive stipends ranging from Rs. 120 to Rs. 200 per month according to their performance. In May 1955, 16 students of the first-year class and 16 students of the second-year completed training. For the new sessions which started in June 1955, 47 students were admitted to the first-year and 20 students to second-year class.

(b) *Short-term statistician's course* : This new course has been arranged to suit persons already in employment, the classes being held in the evening. The course is generally non-mathematical in character with emphasis on the logic and proper use of statistical tools. During the year two courses were arranged with a total of 87 students.

(c) *Officer's course (jointly with CSO)* : This is also a new course, started in September 1955, for the purpose of training officers of the State and Central Governments and other recognized institutions, in different fields of application of statistics. During the year 19 officers took a six months' course, 10 of them continuing for an additional three months specialization course.

(d) *Officers on deputation* : During the year 6 officers were accepted for individual short duration training in different subjects. Besides, 20 other officers were deputed by various institutions to attend Prof. Wiener's seminars.

(e) *Computer's Training Course* : The computers' training course was started in 1951 for the benefit of those desiring to sit for the Computer's Certificate Examination of the Institute. There were two sessions during the year in which 97 candidates participated. [The list of trainees for the various courses is given in Appendix 10].

## TWENTYFOURTH ANNUAL REPORT : 1955-56

### 2. INTERNATIONAL STATISTICAL EDUCATION CENTRE

1. The associated institution, International Statistical Education Centre (ISEC), Calcutta, was opened in October 1950. This Centre is maintained jointly by the International Statistical Institute and the Indian Statistical Institute with the support of the UNESCO and the Government of India. The Centre provides courses of training in theoretical and applied statistics at various levels to trainees from countries of the Middle, South and Far East.
2. From the inception of the Centre upto April 1955, the Centre had conducted eight terms of training, of either six months' or nine months' duration. During these eight terms training was imparted to 191 participants, involving 237 'student terms', and representing 16 Asian countries (Afghanistan, Burma, Cambodia, Ceylon, India, Indonesia, Iran, Iraq, Japan, Malaya, Nepal, Pakistan, Philippines, Syria, Thailand and Vietnam).
3. The Eighth Term, with 24 students from 7 countries, an account of which was given in last year's annual report, closed in the middle of April 1955. The Ninth Term opened on 15 August 1955 and closed on 14 April 1956.
4. The number of trainees admitted to the Ninth Term was 11, distributed among 3 countries as follows : Philippines 3, Pakistan 3 and India 5. Besides these 11 trainees, two trainees from Iran who joined the Eighth Term stayed on for specialization courses till December 1955. Participation during this term was limited in contrast to the previous terms, owing chiefly to the delay in issuing the announcement for the term. Arrangements have been made for making the announcement sufficiently early for the Tenth Term to start in July 1956. Increased participation is expected during the Tenth Term.
5. *Instruction* : Training during the Ninth Term comprised lectures, laboratory work, assisted reading, seminar discussions, in-service training (both at the Indian Statistical Institute and at the Central Statistical Organization (New Delhi) and field work at Giridih, relating to Sample Surveys. The first 3 months of the term were devoted to theoretical work involving about 180 lectures and 350 hours of laboratory work. About three weeks were spent on field work at Giridih and another month on training in official statistics at the CSO. The remaining period had been set apart for training in statistical projects at the Indian Statistical Institute and for specialization courses. The bulk of the lectures were delivered by members of the staff of the Research and Training School. A number of statisticians from different ministries and departments of the Government of India delivered lectures during the training organized by the CSO at New Delhi.
6. *Visiting Professors* : A number of visiting professors arranged for by the International Statistical Institute or the Indian Statistical Institute, or deputed by the UN, and the specialized agencies of the UN have been associated with the instruction at the Centre. The visiting professors during the Ninth Term were Dr. M. Ziauddin (Pakistan), Dr. H. Lubell (USA), Mr. C. A. Links (Netherlands), Dr. P. V. Sukhatme (FAO), Mr. C. K. Dilwali (UN), Dr. W. R. Leonard (UN), Prof. John Maclean (Bombay), Dr. K. S. Banerjee (West Bengal), Dr. T. Podea (USA), and Dr. Q. M. Hussain (Pakistan).
7. The students of the ISEC also attended lectures and seminars given at the Indian Statistical Institute by various distinguished visiting scientists.
8. *Fellowships* : Out of Government of India's contributions to the Technical Co-operation scheme (Colombo Plan), a number of fellowships were being granted to the



trainees of the Centre. During the first eight terms a total of 96 fellowship awards were made. In the Ninth Term six trainees were awarded fellowships.

9. *Certificates* : The 11 students of the Ninth Term who had satisfactorily completed the course received certificates of training at a function held on 5 April 1956.

10. *Student activities* : The ISEC students' association arranged excursions to places in Calcutta, Agra and Delhi. A Souvenir volume for the Ninth Term has been prepared.\*

### 3. PROFESSIONAL EXAMINATIONS

1. *The Statistician's Diploma Examination of the Institute* was held once in August 1955 and again in March 1956 at Bombay, Calcutta, Delhi, Lucknow, Madras and Poona. In all 204 candidates registered for the examination in one or more papers of whom 133 appeared and 77 passed.

2. *The Computer's Certificate Examination* was also held two times during the year (August 1955 and February 1956) at Calcutta, Delhi, Giridih and Poona. 448 candidates registered, 402 appeared and 219 passed.

3. *The Statistical Field Survey Certificate Examinations* (Junior and Senior) were held in August 1955 and February 1956, at Bangalore, Calcutta, Delhi, Lucknow and Poona. 337 candidates registered, 249 appeared and 115 passed.

The names of successful candidates in the different examinations are given in Appendix 12.†

---

\* *ISEC Tenth Term*: The term started on 16 July 1956 and would continue till April 1957. In response to the announcement of the tenth term, 38 applications were received from 8 countries; 30 candidates were admitted; and 29 students from eight countries had joined, the distribution being as follows: Burma-2, Ceylon-2, India-11, Japan-4, Pakistan-4, Philippines-2, Singapore-2, Thailand-2. Training during the term would comprise lectures, laboratory work and assisted reading in methodology, seminar discussions, in-service training at CSO and project work (including field work) at the Institute. An added feature of this term's training is the introduction of 'take-home' periodical tests, which give the students ample opportunities to discuss and understand the subjects better.

† The three examinations were also held in September 1956 at the above centres; 121 candidates registered and 92 appeared for the Statistician's Diploma Examination, 248 registered and 220 appeared for the Computer's Certificate Examination and 62 registered and 47 appeared for the Field Survey Certificate Examination.

## TWENTYFOURTH ANNUAL REPORT : 1955-56

### 4. PROJECT WORK

#### 4.1. National Sample Survey and Associated Projects

1. The statistical work relating to the countrywide continuing National Sample Survey (NSS) has been the charge of the Institute since 1950-51 when the Survey was started. The Field Branch of the Survey has been under the direct control of the Ministry of Finance, except for a special survey unit under the Institute. By the beginning of the year under review, 8 rounds of the survey had already been completed. The 9th round of the survey was started and completed during the year under review, while the 10th round survey work was started and was still continuing at the end of March 1956. On the basis of the analysis of data collected during different rounds, various reports were prepared, and among these, the NSS Report No. 7 : Couple Fertility, and NSS Report No. 8 : Report on Preliminary Survey of Urban Unemployment, were published.

2. In January 1956 the NSS Statistical Section was reorganized into 3 functional units :—(a) Project, (b) Pilot and Research, (c) Special Unit. The functions allotted to the Project Unit were sample selection, preparation of schedules and instructions, clarification of technical points raised by field staff during the course of the survey, tabulation and preparation of summary of results, statistical analysis, drafting and submission of reports and project training. The Pilot and Research unit was charged with improvement of sample design and the assessment of the quality of material, exploring profitable lines of tabulation and analysis etc., and the functions of the Special Unit were project development studies.

3. In the 8th round of the survey the main emphasis was on land holdings with particular reference to operational holding, whereas in the 9th round, emphasis was shifted to collection of information on employment and unemployment. Over and above this, the usual enquiries relating to household consumer expenditure and productive enterprises of household, prices etc., were conducted. Another special feature of this round was collection of data on the small and household industries from households who reported *self-management* under the means of livelihood 'manufacture'.

4. The state of employment, on which special emphasis was laid in this round, was examined from different angles, namely, (a) usual features without particular reference to any short period or point of time, (b) specific features obtaining on a single day of reference, and (c) special features obtaining on each of the seven days of the week of reference.

5. Due to the above changes in the scope of the survey in this round, the sample size was increased but it was considered expedient to try to increase the sample size gradually so that the work of recruitment, training and management of the additional investigation staff as well as of the statistical work was kept within manageable limits. In actual fact there was no noticeable increase in the sample size in the rural sector but in the urban sector it was quite appreciable.

6. There was no substantial change in the scope of the survey conducted in the 10th round except that the collection of data on yields by direct observation by crop-cutting experiments was an additional feature. Emphasis was laid on collection of data on land utilization which included, besides, cultivation of crops, any type of use or non-use of a piece of land. Considerable attention was also given to the collection of data on employment. Questionnaire on village statistics was reintroduced in this round with some modi-



fication with a view to getting a picture of rural life in the country in its various aspects, e.g., marketing facilities, proximity to important communication and administrative points, condition of roads and modes of conveyance, important local crops and sowing practices, educational facilities with their standards, sources of finance and the number of scholars and teachers, medical facilities etc.

7. A training conference in which the field supervising staff of the NSS received their instruction from the technical experts of the Institute as to how to collect the data properly on the schedules and in which the objects, scope and definition of terms used were explained, was held in the Institute premises from 18 to 28 April 1955 and again from 21 to 28 November 1955, the first one in connection with the 9th round of the survey and the second one for the 10th round of the survey.

8. The National Sample Survey primarily meant for estimates of national level was intensified within certain States for the purpose of increasing the reliability of the survey estimates on State levels. This was made possible by the States of UP, Bombay and Travancore-Cochin participating with NSS on the basis of identical coverage and number of sample units. The number of sample units was thus doubled in so far as these States were concerned.

9. An intensive training on the various aspects of project work including training in field survey work and mechanical tabulation was imparted to 19 officers deputed from the different States and the Centre, 11 trainees of the International Statistical Education Centre, and 18 second-year post-graduate students of the Indian Statistical Institute. The training was for an approximate period of 6 months.

10. *Sample survey of manufacturing industries* : As in the previous years, this survey was continuing during the year under review. The reference period of this survey was 1954; field work which commenced from the 1st week of October 1955 and was expected to continue up to the end of September 1956. The special feature of this round of survey was the inclusion of undertakings relating to certain industries included in schedule (1) of the Industries (Development and Regulation) Act of 1951.

#### 4.2. Other Projects and Applied Studies

1. *Socio-Economic Surveys in Rural Areas, Giridih* : Apart from Family Budget Enquiry in Rural Areas which started in June, 1953 the following surveys were conducted during the year under review : Employment Survey, Nutrition Survey, and a survey of Bullock Utilization (Daily) in Rural Areas.

2. *Socio-Economic Surveys in Urban Areas, Giridih* : Family Budget and Employment Surveys were conducted during the period under review and figures of purchases made in workers' families were collected.

3. *Crop-cutting Experiments, Giridih* : Some designs of crop-cutting experiments were tried by the ISEC students and the Institute trainees during their stay at Giridih in November and December, 1955.

4. *Progressive Harvest Surveys, Giridih* : An experiment was conducted during November and December, 1955 with a total of 1125 samples cut in 5 plots.

5. *Field Trials and Precision Studies, Giridih* : Experiments on *Aus* and *Aman* paddy for studying the effect of different methods of sowing, inter-culture, different kinds

## TWENTYFOURTH ANNUAL REPORT : 1955-56

of manures and fertilizers, different spacing and number of seedlings per hole etc., as also the effect of manures on the yield of different methods of plantation continued during the period.

6. *Data regarding rainfall and maximum and minimum temperatures, Giridih* : The collection of these data which had started in 1953 continued during the period under review.

7. *Special Demography, Health and Employment Surveys* : (March, 1955 to October, 1955) : A special Demography, Health and Employment Survey was conducted in 112 urban blocks and villages throughout the State of West Bengal.

8. *Harvest Survey* (October, 1955-March, 1956) : Harvest Survey on a new design with special treatment on the border plants was conducted on different crops. A study of consumption pattern was also attempted for the pre-harvest and post-harvest periods. This intensive study was made in 14 villages of all the districts of West Bengal.

9. *Enumeration of casualties due to cancer for a year in Calcutta* : This study started in August 1955 and it was found that there were about 800 casualties in a year due to cancer in Calcutta.

10. *Pilot Survey on production and utilization of cattle dung* : A short pilot survey was taken up to study production and utilisation of cattle dung in 3 villages of 3 police stations of 24-Parganas, West Bengal.

### 5. ELECTRONIC COMPUTERS

1. Preparatory work for the building up of a well-equipped modern electronic laboratory has been going on for the last three or four years. There was some significant progress this year. An Electronic Computer HEC-2M was purchased from the UK, and valuable electronic equipment also began to be received from the USSR through the United Nations. Two members of the staff, Shri S. K. Mitra and Shri D. S. Kamat were deputed to the USSR in September 1955 by the UNTAA on a special fellowship, to make a technical report on the electronic computer which the USSR Government had offered to the Institute through the UNTAA. They visited Moscow, Penza and other places, where they saw factories and technological institutes concerned with computing machinery. Shri Mohimohan Mookerji and Shri Amaresh Roy also completed their training in the British Tabulating Machines Works at Letchworth, and visited different computing machine laboratories in Europe and returned to India early in 1956.

2. The Hollerith Electronic Computer HEC-2M, was received in February, 1956. The Machine has been installed in an air-conditioned room situated on the ground floor of the Institute building at 203, Barrackpore Trunk Road, and was ready for operation by the end of March, 1956. The Research and Training School of the Institute has many problems to be solved on the machine. Requests have been received from several scientific institutions in India for assistance in computational problems. Efforts are also being made to use this machine for certain classes of data-processing work of the Project Division.

3. A good number of Soviet electronic instruments obtained through the UNTAA have also been received and included six Q-meters and six impulse generators of a very high quality. The remaining items in the list of electronic instruments and the punched card machines to be received from USSR are already on the way. An electronic computer, called the "URAL", much larger in capacity and with greater flexibility than the HEC-2M machine, is expected to be received by the end of 1956.



4. Attempts are being made to organize a computing centre for the benefit of scientific institutions in India. Some service is already being rendered. A problem for preparing a table of values of an improper, convergent integral was received from the Indian Institute of Technology, Kharagpur at the end of 1955, and was solved partly by an electronic calculator. Other problems have been received from the Indian Association for the Cultivation of Science, Calcutta, and from the Indian Institute of Science, Bangalore, which are now being run on the HEC-2M machine.

5. The Analogue Linear Equation Solving Machine, designed and constructed in the Institute in 1953 was reassembled with improvements to achieve greater accuracy. Two electronic Random Number Generators designed, on a new principle, have been constructed and successfully operated.

6. The plan of work on the development of an electronic digital computer, which had began early in 1954, was modified during the course of the year in view of the availability or near availability of machines like the HEC-2M and URAL. The earlier work on a magnetic drum device and associated auxiliary equipment is being incorporated into the design of faster switching and selector circuits.

7. Dr. Clarence Ross, a computing machine specialist in USA spent a few days in the laboratory in November 1955 and had discussions with the staff about programming methods. Messrs. Konoplev and Sychev, engineers, arrived from the USSR in January 1956, to install the punched card machines from the USSR.

8. A number of tape recording machines, speech amplifiers, projectors and other electronic instruments were maintained and repaired by the laboratory as usual. Two speech amplifiers were constructed for the Delhi and Giridih offices of the Institute during the year.

## 6. STATISTICAL QUALITY CONTROL

1. Three SQC Units were in operation during the year at Bombay (started in 1953) and at Bangalore and Calcutta (started in 1954). The activities of these SQC Units cover (a) installation and maintenance of SQC in factories, (b) training for SQC, and (c) promotion of SQC. Service, which is partly subsidized, is given to member firms on a fee basis. At the end of March 1956 there were 32 firm members on the roll.

2. *SQC Policy Advisory Committee:* The SQC Policy Advisory Committee which had been set up in 1954 with Shri C. D. Deshmukh as Chairman and Shri K. C. Neogy (*Planning Commission*), Sir Shri Ram, Shri Kasturbhai Lalbhai, Dr. S. S. Bhatnagar (*Director, Council of Scientific and Industrial Research*), Dr. Lal C. Varman (*Director, Indian Standards Institute*), Shri H. H. Keil, Shri M. G. Kotibhaskar, Shri S. C. Jain and Professor P. C. Mahalanobis as members and Shri Pitambar Pant as Secretary had one meeting on 7 September 1955. The Committee suffered a great loss in the death of Dr. S. S. Bhatnagar. The Committee requested Professor P. C. Mahalanobis to prepare a scheme designed to meet the growing requirements of trained SQC statisticians in the country. The Committee was of the view that the existing scheme of contribution from industry should be continued for the present. The members thought that it would be very useful if the SQC Units could collect material for the assessment of gains arising from the use of SQC and associated methods.

## TWENTYFOURTH ANNUAL REPORT : 1955-56

3. The Committee expressed satisfaction at the decision of the Ministry of Defence to set up a Statistical Quality Control Unit for the Ordnance Factories which would be guided in policy matters by the SQC Policy Advisory Committee.

4. The Committee decided to invite Dr. Thacker, (who had succeeded Dr. Bhatnagar as Director, CSIR) and Shri G. D. Somani, M.P. to join the SQC Policy Advisory Committee.

5. *SQC Unit, Bombay:* The Bombay Unit which was started under Dr. (Miss) S. P. Vaswani in December 1953, made good progress during the period under review. She was assisted by a team of 4 technical officers and a staff of 7. Recently the Unit has been allotted 4000 sq.ft. of space in a new building on Dinshaw Wacha Road. At the end of last year, 9 factories were enrolled as annual members of the Unit. During this year 5 more firms got enrolled and 5 other applications are on the waiting list pending expansion of the Unit.

6. A five-day course in SQC for industrial management was organized by the Unit in Bombay in October, 1955. Admission to this course was restricted to managing directors, managing agents, technical directors and managers of mills and factories. The course was attended by 43 participants from 27 factories including 8 out-station participants from Ahmednagar, Hathras, Sholapur, Ratlam, Sidhpur, Moradabad and Ellichpur. This was followed by a 2-month training course for technicians; 54 trainees including 6 from outside Bombay, nominated by 30 factories representing textiles, automobiles, engineering, cycles, iron and steel, ordnance, oil, matches and silk industries were admitted to the course. In-factory training for inspectors, operative and setters were also organized in member mills yielding good results. Also, seven apprentices and four officers deputed by institutions were given training with the SQC Unit. Lectures, talks and discussions were also organized at various institutions and promotional work was done and studies made in several factories. Visits were also paid to 6 factories on request from management.

7. *SQC Unit, Bangalore:* Shri R. Natarajan was the Administrative Officer of the Unit till the middle of September, 1955; and was succeeded by Shri A. K. Ghose, ICS, Managing Director, Bharat Electronics Ltd., Mr. T. Hanada, SQC expert from Japan, worked with the Bangalore Unit from December, 1955 till the end of March, 1956. There were, on 31 March 1956, 10 establishments on the roll as members of the Unit.

8. A number of lectures and talks and discussions were arranged by the Unit at institutions and societies concerned with industry. Some talks were also arranged in co-operation with the Quality Control Association, Bangalore, and papers were submitted by the Unit for discussion at the Chemical & Textile Standards Convention of the Indian Standards Institute held at Bombay from 9th to 15th January, 1956.

9. At the meeting of the Advisory Panel of the Unit held on 7th and 8th May, 1955, in Bangalore, Prof. S. K. Ekambaram gave a talk on "SQC in Industry" and Messrs. E. H. Osman and A. R. Frederick of the US Technical Co-operation Mission to India spoke on "The Role of Standards in Quality Control", and "Managerial Prerequisites of Quality Control" respectively. Practical demonstration of SQC methods and techniques were given by the Unit staff from materials brought from the production lines in factories. Visits to 19 factories and mills were undertaken by the staff of the Unit and guidance in technical work was given.



10. A 10-Day course was conducted between 9th and 19th May, 1955 in Bangalore by the Quality Control Association in which the Unit staff fully participated. A second similar course for managers and supervisors was held in Coimbatore from 19th to 29th December, 1955 by the Unit. This was attended by 25 candidates from textile, rubber, textile-machinery, engineering and other industries.

11. *SQC Unit, Calcutta*: The Calcutta Unit which had been established in September 1954 got into stride during the year under review. The staff consisted of 6 technical assistants and 2 others at the end of March 1956. The Unit was fortunate in having technical guidance from Messrs. G. Taguchi, and T. Hanada from Japan and Mr. D. J. Desmond from the United Kingdom. Mr. Taguchi, who was associated with the Unit from the beginning returned to Japan in the middle of August 1955. Mr. Hanada worked with the Unit from April to December 1955. Mr. Desmond, whose services had been made available by the Government of UK under the Colombo Plan, was in charge of the Unit for the rest of the period under review.

12. The Unit started its regular service to industry in April 1955 with three factories as members. By 31 March, 1956, 5 more firms had joined, bringing the total number of members to 8. The Unit arranged for two courses for training factory-technicians in SQC methods. A course of lectures on Design of Experiments was given by Mr. Taguchi to the technicians of the Indian Aluminium Co. Ltd., Belur, in April 1955. An intensive course of ten working afternoons, was organized from 17 to 31 May, 1955 for the SQC technicians of the member factories. Facilities were also provided by the Unit for practical training in SQC for officers from outside.

## 7. WORK ON PLANNING FOR NATIONAL DEVELOPMENT

### 7.1. Economic Planning

1. A small Operational Research Unit (ORU) for planning was established early in 1954, and work was started on a small scale. Prime Minister Jawaharlal Nehru inaugurated studies on planning for national development in the Institute on 3 November 1954. Work on planning was then organized on a large scale and culminated in the preparation (by collaboration between the Institute, the Central Statistical Organization, the Ministry of Finance and the Planning Commission) of the Draft Plan-frame of March 1955. This was accepted by the National Development Council (which consists of the Central Cabinet, the Planning Commission and the Chief Ministers of all the states of India) as the basis for the formulation of the Second Five Year Plan of India.

2. In September 1955 it was decided by Government that the technical and statistical work on the Second Plan in the Institute should continue and greater attention should be paid to perspective planning and more especially to statistical work in this connexion. It was further agreed that in order to facilitate this work the operational research unit (ORU) in the Institute should continue on an expanded scale in close collaboration with the Central Statistical Organization, the Ministry of Finance, and the Planning Commission. In accordance with the above decisions by Government, the staff for work on planning was expanded and an economic wing was set up in the Institute in the last quarter of the financial year 1955-56. Since then work on planning in the Institute has been continuing in both Calcutta and Delhi and is organized in several broad groups.

## TWENTYFOURTH ANNUAL REPORT : 1955-56

3. One section was engaged in studies on methodological aspects of long-term planning. Professor Charles Bettelheim of Paris, who came to the Institute in October 1955 with an assignment under the United Nations, was actively associated with this work. The general approach in this section is to assess physical requirements and available resources; and to work out plans for various periods with the help of input coefficients and technological constants. Work was also started on transport requirements, and on the coal industry in all its aspects.

4. Another section continued the study of inter-industry relations which had been started in 1954 on the lines of Leontief's work. A table of inter-industrial transactions for 1951-52 was completed which distinguished between 34 broad groups of activities, and showed the utilization of the product of each activity in the form of input into one or more of the other activities as well as in private and public consumption, capital formation and export. The construction of another table of transactions for 1953-54 was in progress. Analytic studies had also been started on estimating levels of activity appropriate to any assigned target of final demand: and on forecasting the final demand in 1953-54 on the basis of the coefficients of 1951-52 for purposes of comparison.

5. A third group was engaged on economic studies of a more general nature covering such topics as problems of differential excise duties, interaction between the monetized and the non-monetized sectors of the economy, labour productivity in Indian manufacturing industries, capital-output ratios, monopoly concentration in Indian industries, concepts of national income appropriate to underdeveloped countries etc. A report on a special survey of the Chittaranjan township was in progress.

6. In Delhi the Institute staff worked in close collaboration with the Central Statistical Organization (CSO) and undertook a number of special studies among which may be mentioned the problems of financing the Second Five Year Plan, and the statistical relationship between changes in national income and changes in the production of basic commodities like steel, coal, electricity and cement in the UK, USA and the USSR.

7. Besides studies on planning, the Planning Division collaborated with the National Sample Survey for the framing of a pilot schedule for the agricultural labour enquiry and the price schedule of the eleventh round of the NSS.

8. The work on planning greatly benefited from discussions with foreign scientists who gave lectures and participated in seminar work in the Institute. Special mention may be made in this connexion of Professors, Paul Baran (USA), V. Dyachenko (USSR), Charles Bettelheim (France), J. K. Galbraith (USA), Oskar Lange (Poland), J. A. Links (Netherlands), The Rt. Hon'ble John Strachey, M. P. (UK), J. Tinbergen (Netherlands). Professor Norbert Wiener (USA) set out in a seminar lecture the theoretical considerations involved in work on planning. Dr. Frank Yates (UK) gave some preliminary considerations to agricultural planning.

9. Dr. Nicholas Kaldor of the University of Cambridge, UK, came to India at the invitation of the Indian Statistical Institute. By special arrangement he worked for nearly three months in the Ministry of Finance in New Delhi and prepared an important Report on Indian Tax Reform. He also gave lectures and had discussions in Delhi and Calcutta.

10. *Working Papers on Planning*: A series of working papers on planning, which had begun to be issued from November 1954, were continued and 22 Working Papers were released during the year under report. A list is given in Appendix 9.



## 7.2. Kalyanashree

A new unit called *Kalyanashree* was started in Calcutta on 28 February, 1956 to collect statistics and study the economics of small-scale household industries and handicrafts. The Institute has supplied accommodation and some equipment and also provides staff for the statistical and economic studies. The actual work in the unit is organized on an entirely self-supporting basis and current expenses are met from sales of production. Extracts from the speech delivered by Professor P. C. Mahalanobis at the opening ceremony are given below.

"1. I shall try to explain very briefly the purpose which we have in view in starting this new unit. There are no immediate prospects of our being able to get rid of unemployment exclusively through modern industrial developments. We must, of course, develop modern industries as fast as we can; but it would take a number of years, 5, 10, 15 or even 20 years before we can solve the problem of unemployment in the way it was solved in highly industrialized countries of the world. In this situation, as we have large resources of raw materials, our thinking in India is turning to the possibility of using our idle hands to increase production in the country as much as possible. Such production, we are aware, would not be immediately as efficient as production with the help of modern machinery driven by power. But it is not merely the production per hour or per person which is important—it is the total production in the country which matters; and, therefore, in India we are thinking seriously of activating the unexploited resources which are not being utilized at present; and the biggest pool of such resources is, of course, the idle hands. Therefore, we are eager to know what are the possibilities of increasing production in household enterprises.

"2. In this situation, in our own Institute, we are thinking of making a thorough study of the problem through partly controlled laboratory-type experimentation and also through field studies. As a first step, we are establishing this small centre for traditional handicrafts where those who have no employment would come and would be given tools and raw materials with which they would produce what they can. And we shall give a guarantee to purchase what they would produce. We shall, most of all, draw upon skills which are already there but are not being used. We would also supplement this, to some extent, by giving training to the unemployed (many of whom would be refugees). The aim would be to help them to earn their living as quickly as possible.

"3. \* \* \* Our statisticians, our economists, and our technologists and scientists would use this centre as a laboratory to study what would be the output, what would be the cost, and what would be the effort needed to organize the work and so on.

"4. This is the first step. It is our hope that very soon the work would be extended to some of the households in our neighbourhood. We shall try to supply raw materials, perhaps also some small tools to people who will work in their own homes—that would be our second step. We may then extend the study wider. We may take a group of 4 or 5 villages about 3 or 4 miles away but still within easy reach so that contact can be maintained; and we shall supply raw materials or tools and try to increase production in the village enterprises. This would be the third step. We have also a fourth step in view, namely, to establish a small independent centre with perhaps one worker or may be with two for a group of 8 or 10 or even 20 villages. In these villages efforts would be made to increase production by using idle hands and skills which are already there.

## TWENTYFOURTH ANNUAL REPORT : 1955-56

"5. At the same time, our economists and statisticians would study the effect of an increase of production and income. It will be the task of the National Sample Survey, the Special Technical Unit, and the Planning Division, to find out how the unemployed are living before they start work in our centres. Our survey staff will go to their home and find out what is their income, what they are now eating, and how they are living. When they begin to work in our centre and their income goes up, we shall try to find out how that additional income is being spent, and what are the new demands. This we hope to do not only in the indoor centre but in the outdoor unit; and, later on, also in the village unit. This, we hope, will throw some light on the question of what would be the new demands which would be created by the new purchasing power.

"6. There is a third aspect which also we are keeping in mind, namely, the evaluation of the product at market prices, and also in relation to what should be considered some kind of a living wage. That is, we want to find out to what extent subsidies would be required and what would be the amount of such subsidies, to enable the products being sold in the market; of course, we may find that in certain lines no subsidies are needed. We shall also try to learn something about the marketing aspect of the problem. Finally, we are keeping in mind the need for improving the tools for household enterprises.

"7. \*\*\* In our household enterprises in which we have a tradition of welfare wedded to beauty—the name which has been selected for the new centre gives expression to this idea—*Kalyana* may be, in a general way, translated as "welfare", and *Shree* is "beauty and grace". It is a great tradition of India to look upon production not merely from the aspect of value but also from the aspect of welfare and the aspect of beauty. Therefore, I think *Kalyanashree* which is welfare and beauty wedded together is an appropriate name for the new unit."

### 7.3. Industrial Management Research Unit for Planning (IMRUP)

The Council of the Institute approved in March 1956 a scheme for the setting up of another special unit with headquarters at Bangalore to study the problems of industrial management in both the public and private sectors. Two experienced engineers of senior standing, Dr. Bhola D. Panth and Shri D. P. Basu, started preliminary work in consultation with another senior engineer, Shri R. Natarajan who joined the new Unit in the Institute a little later in May 1956.\*

---

\* The work of the IMRUP began with a preliminary Symposium on "Organization and Management of Public Enterprises" held in Bangalore on 17 and 18 April 1956 which was inaugurated by S. C. Sen, Joint Secretary of the Institute, and was attended by R. Natarajan as Chairman, and D. P. Basu (Director, IMRUP; Director ACBI Ltd.), M. N. Dastur (Industrial Consultant, Managing Director, M.N. Dastur & Co Ltd.), D. J. Desmond (Adviser on SQC under the Colombo Plan), J. K. Galbraith (Professor, Harvard University, USA), A. K. Ghosh (Managing Director, Bharat Electronics Ltd.), P. K. Gopalakrishnan (Research Associate, IMRUP), S. S. Khera (Secretary, Ministry of Production), K. B. Madhava (Actuarial Consultant), M. K. Mathulla (Managing Director, Hindusthan Machine Tools Ltd.), K. Narayanaswamy (Director of Industry & Commerce, Mysore State), Pitambar Pant (Secretary to the Chairman, Planning Commission), Bhola D. Panth (Director, IMRUP), T. Shamanna (Vice-Chairman, Mysore Iron & Steel Works Limited), J. M. Shrinagesh (Managing Director, Hindusthan Aircraft Ltd.), the Rt. Hon'ble John Strachey, (Member, British Parliament), and P. C. Suri (Director, Public Management Studies, Planning Commission). The proceedings of the Conference have been circulated separately.



## 8. SERVICE UNITS

## 8.1. Machine Tabulation Unit

1. The Machine Tabulation Unit was responsible primarily for (1) preparation, (2) processing and (3) storing of punched cards relating to statistical work of National Sample Survey and allied projects undertaken by the Institute. This Service Unit accommodated also special types of calculation work of the research workers of the Institute. In the year 1954-55 the size of the Machine Tabulation Section doubled at Baranagar and this necessitated, early in 1955-56, experienced workers from other organizations to be requisitioned to meet up the personnel deficiency. By the latter part of the year a new tabulating unit started functioning at Delhi. An experienced senior worker from Baranagar was sent to Delhi to organize the tabulation unit there solely for NSS work. The strength of tabulation equipment at Baranagar, Giridih and Delhi, as they were in December 1955 and in March 1956 are given in Appendix 7.

2. The rapid expansion of this Section created a number of problems of which scarcity of trained operators was the foremost. A batch of young college boys were recruited and trained in punched card system. At the end of training 11 candidates turned out successful and were absorbed as machine operators at Baranagar.

## 8.2. Library

1. The Central Library was located at Baranagar with service branches at the City Office at 9/B, Esplanade East, Calcutta, and at the Giridih Branch of the Institute.

2. The Library acquired 2717 volumes of books and received 1190 periodicals and annuals against 1945 volumes of books and 1126 periodicals and annuals last year. The total number of books, journals and other materials issued was 38,055 against 23,033 last year.

3. Amongst the various services rendered by the Library may be mentioned (1) Bibliographical Services, (2) Circulating Library, (3) News Clippings—a new unit started in June 1955 for systematic processing and indexing of relevant material relating to statistics and planning in India, and (4) Translation: letters, articles, books in other foreign languages specially in French and Russian were translated into English.

4. The Records Unit under the Library Service functioned as usual, classifying and indexing files, maintaining and preserving schedules, working papers and reports relating to surveys and projects carried out in the Institute, processing and shelving through serial sorting and numbering cadastral survey maps of the districts of West Bengal and the State maps of the Indian Union. [Some further details are given in Appendix 8].

## 8.3. Workshops

1. The workshops of the Institute began to function on an expanded scale during the year. The development sector was responsible for the repair, maintenance and development work of the Electronic Computer section and catered for other research needs. It made considerable progress in improving the proto-type desk calculating machine which was designed in 1954-55 and repaired 671 desk calculators of different types. A number of additional equipment were installed in this sector during the year



## TWENTYFOURTH ANNUAL REPORT : 1955-56

which made it necessary to expand the floor space from about 1000 sq. feet to nearly 4000 sq. feet. The number of workers also increased from 19 to 37.

2. The general sector of the workshop, which had 12 workers including some highly skilled mechanics, looked after the general requirements of the Institute. It constructed one heavy duty handpress (capacity 40 tons), one grinding machine, one blade grinder and several other small tools. This sector manufactured 250 pieces of crop-cutting apparatus and six sets of SQC models with improvements devised in the workshops. It also executed a large number of maintenance jobs.

3. The project for securing machine tools from the USSR through the United Nations Technical Assistance Administration materialized during the year with the arrival of the first consignments of Soviet machine tools. Some of these machine tools are being installed in the workshops.

### 9. SOCIAL AND WELFARE SECTION

1. *Health Home:* The foundation stone of the Health Home for workers was laid by Dr. Satyasakha Maitra, F.R.C. S. at Giridih on 26 December 1955; and the Home was formally opened by Dr. F. Yates, F.R.S., on 5 March 1956. Professor P. C. Mahalanobis presided over the opening ceremony at which Mr. E. A. Rowse, Shri C. V. Narasimhan, (Joint Secretary, Ministry of Finance) and Shri N. S. Pandey (Deputy Secretary) were present among the guests. The Health Home was built on land received as a gift from Sm. Nirmal Kumari Mahalanobis with an initial donation of Rs. 5150/- received from the delegation of Soviet scientists in 1954-55.

2. *Medical Welfare Unit:* The Medical Unit at headquarters, under the charge of Dr. R. Maitra, M. B., has made steady progress and is at present housed in an independent building with a well-equipped dispensary, a reception room for patients, an examination room, an isolation room and two single-seated sick rooms.

The benefits of the Unit cover the workers and their families and include free treatment at the attached dispensary, calls at Workers' and Officers' residences by the medical officers at the subsidized rate of one rupee and eight annas per call, and prescriptions supplied on a cost basis.

The total number of cases treated during the year under report was 8,819 against 7,449 during 1954-55. The Medical Officer attended 830 calls at workers' residences, the corresponding figure for the previous year being 740. Medical consultations, minor surgical treatment and injections etc., totalled 757 against 688 in the preceding year. In all 9,079 prescriptions were served by the Unit Dispensary, the corresponding number for the previous year being 7,214. Medicine worth Rs. 10,332 was supplied to workers during the year; the cost on this account during the previous year was Rs. 7,877 only.

Six-monthly inoculations against typhoid and cholera and yearly vaccination against small pox were given as usual to the entire Institute staff at Headquarters and City offices. The benefit of the anti-malarial measures, which was so far confined to the Institute workers only, has been extended to cover the entire Field staff. The Medical Unit also looked after the sanitation of the Institute campus, including hostels and other residential quarters with the active co-operation of the Estate Office. The Medical Unit has started a preliminary statistical study on health conditions of Institute workers.

In the Branch Medical Unit at Giridih under the care of Dr. N. K. Das, M.B., 2073 prescriptions were served and the Medical Officer attended 377 calls from the workers' families. Although there is no regular Unit attached to the ISI Office at New Delhi there exists an arrangement for medical aid to the workers and guests. The services of Dr. B. Saha, M.B. are available to the guests and workers as and when required.

3. *Night School and Adult Literacy Drive:* At the beginning of the Session the Night School had 40 students on its roll. Two new teachers were appointed and a School Committee of 4 members has been formed. The School now has some permanent accommodation and some essential equipment has been obtained.

As a part of the ensuing 25th Anniversary Celebrations a drive against illiteracy has been launched among the workers of the Institute. Fiftyone students have already been enrolled and judging by the progress already made it is expected that all of them will become literate before the Anniversary celebrations in December 1956.

4. *Workers' Club:* During the year under review the Club at Baranagar made substantial progress in every sphere of its activities. The membership registered a further increase and stood at about 600 at the close of the year.

The Sports and Games Section of the Club arranged inter-section football and volleyball tournaments. Competitions in badminton, chess, cards (auction bridge) and table tennis were organized during the year. The fourth annual sports meeting was held in January 1956.

The Social and Cultural Section organized debates, film shows and an exhibition of photographs by Shri B. K. Sinha which was the first of its kind in our Institute. A large number of members participated in recitation competition arranged by the Club and a steamer trip was organized in February 1956. The third annual number of 'Lekhan', the annual literary organ of the Club, won wide appreciation and several issues of a wall paper came out during the year. On the occasion of the fourth social gathering and prize distribution ceremony held in March, 1956, the Club staged "Natun Prabhat" by Shri Manoj Bose.

The Club undertook successfully the formation of a Workers' Co-operative Credit Society. The ISI Workers' Tuberculosis Bed Fund sponsored by the Club was also started with an initial sum of Rs. 3,000/- earned by the voluntary service rendered by a large number of Club members in connection with the field work of the "Workers' Consumption Expenditure Survey" undertaken by the Planning Division of the Institute.

5. *Public Relations Work:* The growing public interest in our Institute has been reflected in an increasing flow of visitors. During the first 3 months of 1956 no less than 150 distinguished visitors came to the Institute. A Reception Room attended by a receptionist was set up in January, 1956. A brochure, explaining the work of the Institute, has also been published.

6. *"Samvadadhvam":* It had been felt for a long time that it would be a good idea to have a house journal that would serve as a link between the different departments of the Institute and provide a common medium for exchange of information and ideas between the entire body of its workers. Plans for bringing out such a journal were taken in hand during the year and the first issue came out under the title "Samvadadhvam" (taken from a Vedic text which means "united we speak"). It is hoped that the journal will fulfil its purpose by fostering a sense of solidarity among the workers.



## TWENTYFOURTH ANNUAL REPORT : 1955-56

7. *Canteen*: Besides the routine serving of meals and refreshments to the workers, the canteen undertook the catering arrangements for several tea and luncheon parties.

8. *Salboni Club*: The Salboni Club at Giridih celebrated the Bengali New Year's Day, the Republic Day, 'Barsha Mangal' and the Anniversary of the Foundation Day of the Institute. It held social functions on different occasions such as the opening of the Club's present premises, inauguration of the Health Home, and reception to the ISEC trainees and RTS students. The Club provided facilities for indoor and outdoor games, staged a play, held an annual sports meeting, and organized tournaments. It also organized for its members a sight seeing and educational tour to the DVC area. The Club raised a fund of Rs. 138/6/- to provide medical relief to an ex-worker.

### 10. EXTERNAL ACTIVITIES

1. Professor P. C. Mahalanobis (accompanied by Mrs. Mahalanobis) and Dr. C. R. Rao attended the biennial session of the International Statistical Institute held in Rio de Janeiro, Brazil, from 24 June to 2 July 1955. Dr. C. R. Rao also attended the Biometric Conference at Campinas, Brazil, in the same month. Professor and Mrs. Mahalanobis visited New York and London on their way back.

2. Professor Mahalanobis attended the meetings of the Working Group of Experts on Family Living Studies in the International Labour Organization (ILO : Geneva, September 1955); Study Group on the Measurement of Levels of Health of the World Health Organization (WHO : Geneva, October 1955).

3. Shri Samar Kumar Mitra and Shri D. S. Kamat visited the USSR in September-October 1955 for discussions on electronic computers.

4. Shri Ajit Das Gupta and Shri Ranjan Som attended the first meeting of the Working Party of the Economic Development and Planning held under the auspices of the ECAFE at Bangkok from 31 October to 12 November 1955 and also attended a seminar on Population Studies held in Bandung in November 1955.

5. Professor Mahalanobis presided over the Third Pakistan Statistical Conference at Lahore in February 1956.

6. Shri Debabrata Lahiri attended the Fourth Regional Conference of Statisticians organized by Economic Commission for Asia and the Far East at Bangkok from 26 March to 6 April 1956.

7. Dr. C. R. Rao attended by invitation the Third All Union Mathematical Conference held in Moscow from 24 June to 5 July 1956.

8. Shri S. Chatterjee, Dr. E. Harper, Shri S. Mitra, Dr. Des Raj, Shri S. P. Sangal Shri Sushil Kumar and Shri T. K. Sen attended the Agra Session of the Indian Science Congress in January 1956.

9. Dr. Shib K. Mitra attended the Third All-India Seminar on Educational and Vocational Guidance held at Baroda in February 1956.

10. Dr. C. R. Rao, Dr. G. Kallianpur, Dr. A. Matthai and Dr. Des Raj served as members of the Board of Studies and also on the Board of Examiners of different universities.



## 11. VISITING PROFESSORS

As in previous years the Institute was fortunate in having a number of distinguished scientists as visitors from France, Israel, Japan, Netherlands, Norway, Pakistan, Poland, Sweden, the UK, the USA and the USSR.

The following scientists came to India at the invitation of the Institute and worked here for several months.

1. PROFESSOR PAUL A. BARAN, *Stanford University, USA* (December 1955 to February 1956). Lectures : (i) National Income Concepts and Computations and (ii) Theory of Imperialism Reconsidered. Discussions on Planning.
2. PROFESSOR CHARLES BETTELHEIM, *University of Paris, France* (Through United Nations : October 1955 to August 1956). Research on Economic Planning.
3. PROFESSOR J. K. GALBRAITH, *Harvard University, USA* (February to April 1956). Lectures : Strategy of Inflation Control. Discussions on planning. Paper: "Economic Planning in India : Five Comments".
4. DR. ARTHUR GEDDES, *University of Edinburgh, UK* (September 1955 to January 1956). Lecture : Traversing India a geographer uses Economic Statistics. Discussions on the use of NSS data for geographical purposes.
5. DR. Q. M. HUSSAIN, *University of Dacca, Pakistan*, (ISEC : September 1955 to February 1956). Lecture : On Design of Experiments.
6. PROFESSOR OSKAR LANGE, *Rector, Institute of Statistics and Planning, Warsaw, Poland* (January to May 1956). Lectures : (i) Organization and Working of Socialistic Economy, (ii) Theory of programming. Discussions on planning.
7. DR. J. A. LINKS, *Central Planning Bureau, Netherlands* (ISEC : September 1955 to February 1956). Lectures and research on planning.
8. MR. H. LUBELL, *Falk Project for Research, Israel* (ISEC : September to October 1956). Lectures on International Accounts, Studies on analysis of consumer expenditure.
9. DR. NICHOLAS KALDOR, *Cambridge University, UK* (January to April 1956). Prepared report on "Indian Tax Reform" (published by the Government of India). Also lectures on Indian taxation.
10. PROFESSOR S. N. ROY, *University of North Carolina, USA* (February to April 1956). Lectures on "Statistical analysis" with seminar discussions on mathematical statistics.
11. THE RT. HON'BLE JOHN STRACHEY, M.P., *UK* (February to April 1956) : Lecture : Contemporary Capitalism and the Under-developed Areas; Discussions on planning.
12. PROFESSOR A. S. SOBOLEV, *Academy of Sciences, USSR* (January 1956) : Lecture: Closure theorems of an algorithm in solving integral equations.
13. DR. G. TAGUCHI, *the Institute of Tele-communication, Tokyo, Japan* (September 1954 to August 1955) : Consultation work in Statistical Quality Control.
14. DR. DANIEL THORNER, *the Institute of Oriental Studies, Philadelphia, USA* (November to December 1955) : Lectures on Agrarian problems in India.
15. DR. J. TINBERGEN, *Planning Bureau, Netherlands* (January to February 1956). Lectures and discussions on Econometric models in planning.
16. PROFESSOR NORBERT WIENER, *Massachusetts Institute of Technology, USA* (September 1955 to April 1956). A course of 60 lectures on (i) Harmonic analysis, (ii) Ergodic theorems, (iii) Prediction problems, single and multiple with seminar discussions.
17. PROFESSOR S. S. WILKS, *University of Princeton, USA* (March to April 1956). Lectures on (i) Ordered statistics, (ii) Variance components, (iii) Government statistics in the USA, with discussions.

## TWENTYFOURTH ANNUAL REPORT : 1955-56

18. DR. FRANK YATES, *Rothamsted Experimental Station, UK* (March to April 1956). Lectures : (i) Programming problems, (ii) Long term experiments with discussions.
19. DR. M. ZIA-UD-DIN, *University of Punjab, Pakistan* (ISEC : August to September 1955). Lectures : (i) Statistical methods, (ii) Symmetric functions.

### FOREIGN LECTURERS

A number of foreign scientists visited the Institute for short periods and kindly gave lectures and had discussions in the Institute.

1. DR. MARCUS BACH, *University of Iowa, USA* (February 1956). Lecture : Mahatma Gandhi & Albert Schweitzer : Links between East and West.
2. MR. C. K. DILWALI, *UN Statistical Office, New York* (January 1956). Lecture : UN Statistical data processing.
3. PROFESSOR V. P. DYACHENKO, *Corresponding Member of the Academy of Sciences, USSR, Institute of Economics, Moscow* (January 1956). Lecture : Organization of Economic Research in the USSR with discussions.
4. DR. J. DOWNIE, *UK* (August 1955). Lecture : Competition theory and its relation to Macro-economics.
5. DR. ERLAND V. HOFSTEN, *Chief of the Statistical Division of the Social Welfare Board, Sweden* (November 1955). Lectures : (i) Household enquiries, (ii) Some new problems in Index numbers.
6. DR. GUNNAR MYRDAL, *Executive Secretary, Economic Commission of Europe, Geneva* (January 1956).
7. PROFESSOR EDWARD NAROZEWSKI, *Poland* (March 1956). Lecture : Problems of measures.
8. DR. TITUS PODEA, *Economic Consultant, New York, USA* (February 1956). Lectures : Planning in USA.
9. PROFESSOR S. M. TUAN and S. CHENG, *China*, Lecture : Some mathematical problems.
10. PROFESSOR S. KRITSKY, *USSR* (January 1956). Lecture : Problems in river research.
11. DR. P. K. WHELPTON, *Director of Scripps Foundation for Research in Population Problems, USA* (January 1956) Lecture : Fertility analysis.

### INDIAN LECTURERS

We are also grateful to a number of Indian colleagues who gave lectures in the Institute. A list is given below.

1. DR. K. S. BANERJEE (*Deputy Director, State Statistical Bureau, West Bengal*) : Lectures on Constructions of Cost of Living Index Numbers.
2. MR. K. C. CHERIYAN (*Agricultural Credit Department, Reserve Bank of India, Bombay*) : Lectures on Rural Credit Surveys.
3. PROFESSOR V. M. DANDEKAR (*Gokhale Institute of Politics and Economics, Poona*) : Lecture on Fundamental concepts in Fisher's theory of estimation.
4. MR. C. R. B. MENON (*Director General of Commercial Intelligence and Statistics, Government of India*) : Lectures on Statistics of trade and industry.
5. PROFESSOR J. MACLEAN (*formerly of Wilson College, Bombay*) : Lectures on Problems in Mathematics.
6. DR. G. B. RAO (*Deputy Director of Commercial Intelligence and Statistics, Government of India*) : Lectures on Statistics of trade and industry.
7. DR. K. S. RAO (*University of Bombay*) : Lectures on Econometric problems.



## 12, GENERAL ADMINISTRATION

1. *Membership*: The position of membership of the Institute during the year under review was as follows:

	new members enrolled during the year	total number at the end of the year
Ordinary members	33	146
Sessional members	nil	9
Student members	6	21
Life members	1	48

2. *Honorary Members*: On the recommendation of the Council the following were elected Honorary Life Members:

Shri Chintaman D. Deshmukh who has been helping the Institute as its President since 1945.

Dr. Satya Churn Law, M. A., B. L., Ph. D., who has been holding the Office of Treasurer since 1936.

Professor K. B. Madhava, A. I. A., who helped in the preparatory work which led to the foundation of the Institute in 1931 and has been associated with its work since then in many capacities.

Sm. Nirmal Kumari Mahalanobis who has been actively helping the Institute since its foundation.

Dr. H. C. Sinha, M. Sc., Ph. D., who had helped in the preparatory work leading to the foundation of the Institute in 1931 and gave effective help as Joint Secretary in the earlier years.

3. *Council*: Names of the members of the Council are given in Appendix 1. The Council held nine meetings during the year. Among the important matters considered by the Council may be mentioned the Rules of the General Provident Fund of the Institute and the Staff Insurance Scheme (20 April 1955); establishment of the Industrial Management Research Unit (20 March 1956); and changes in the Rules of the Institute (24 March 1956).

4. *Governing Body*: The Governing Body of the Research and Training School met at Baranagar on 13 August 1955 with Sir D. N. Mitra in the Chair. The Finance Committee of the Governing Body also met on the same day at Baranagar. (Names of the members of the Governing Body the Finance Committee and other Committees set up by the Council are given in Appendices 2 and 3).

5. *Headquarters*: During the year under review an area of 18,000 sq. ft. was built upon in addition to the existing floor space of 44,000 sq. ft. in the main building at 203, Barrack-pore Trunk Road. Moreover, to meet the demand for more space for project work an increased floor space of 25,000 sq. ft. was made available by erecting asbestos sheds. The carpentry and smithy attached to the Headquarters met, during the year under review, the whole requirement of furniture and Hollerith Cabinets for Baranagar, Giridih and Calcutta offices.



## TWENTYFOURTH ANNUAL REPORT : 1955-56

6. *Calcutta City Office:* The Calcutta Unit of Statistical Quality Control which was established in September 1954 had its office at 9B, Esplanade; this office was also used for general administration work and for meetings and lectures of the Institute. Other offices were located at 294/1, Upper Circular Road, Calcutta-9 and 210, Cornwallis Street, Calcutta-6.

7. *Giridih Office:* During the year under review development work on land progressed satisfactorily. A survey of the entire land was carried out, and a contour map was prepared. A Health Home was opened in February 1956 (further details in Section 9). The students of the professional training course and ISEC trainees went to Giridih for field experiments and training and various experiments and surveys were organized there. (Further details in Section 4.2.)

8. *Delhi Office:* The office at Delhi continued to function at 8, King George Avenue as a link between the Institute and the Central Statistical Organization, the Planning Commission, the Department of Economic Affairs and other departments of the Ministry of Finance and other Ministries.

9. *Bombay Office:* The Bombay Office helped the Branch in the work of ninth and tenth rounds of National Sample Survey in Bombay city as also in conducting Statistician's Diploma Examinations and organizing a series of lectures, the details of which are given under section 13.1. The report of activities of the Statistical Quality Control Unit at Bombay is given under Section 6.

10. *Bangalore Office:* The Bangalore Office helped in the work of two pilot sample surveys, the details of which are given under Section 13.2. The report of activities of Statistical Quality Control Unit at Bangalore is given under Section 6.

11. *Field Branch:* The Field Branch of the Institute had a staff of 161 workers on 31 March, 1956. During the year under review, series of short-term experimental surveys were taken up along with NSS work. Among such surveys may be mentioned enquiries relating to demography, health and employment, harvest crops, casualties due to cancer year in Calcutta, and a pilot survey to study the production and utilization of cattle dung in selected villages etc.

12. *Distribution of staff at different centres:* The following table shows the distribution of workers at Baranagar, Calcutta, Giridih, Delhi, Bombay and Bangalore as on 31 March 1956 as compared with the figures at the end of the previous year. Figures are also shown for the Field Branch maintained under the direct control of the Institute for sample surveys and special enquiries.

centre	general workers		subordinate staff		total	
	1955	1956	1955	1956	1955	1956
Calcutta Headquarters	548	756	161	290	709	1046
Calcutta City	26	9	8	9	34	18
Giridih	70	100	16	29	86	129
Delhi	6	28	3	5	9	33
Bombay SQC	7	10	3	1	10	11
Bangalore SQC	8	8	1	1	9	9
IMRUP (Bangalore)	nil	1	nil	nil	nil	1
Statistical	665	912	192	325	857	1247
Field Branch	120	132	38	41	158	173
total	785	1044	230	376	1015	1420

13. *Changes in Staff:* Some of the senior workers who joined the Institute during the year are mentioned below together with the respective joining date and, within brackets, the name of the Division: Shri Satyasankar Sengupta, 1 May 1955 (*Planning Division*), Dr. Shib Kumar Mitra, 12 July 1955 (*Psychometric Unit*), Shri Sidhartha Banerjee, 19 September 1955 (*Organization and Methods Unit*), Dr. B. C. Das, 20 October 1955 (*Biometric Research Unit*), Dr. Rhea Das, 20 October 1955 (*Psychometric Unit*), Professor Panchanan Chakravorty, December 1955 (*Planning Division*), Shri Mohanlal Ganguli, 2 January 1956 (*National Sample Survey*), Dr. Bhola D. Panth, 3 January 1956 (*Industrial Management Research Unit for Planning*), Professor P. Sarbadhikari, 9 January 1956 (*Planning Division*), Shri Ajit Kumar Biswas, 15 March 1956 (*Planning Division*).

Shri S. B. Sen who had gone on leave in May 1954 to take up an assignment under UNTAA Programme in the Philippines returned and rejoined the Institute to work on a part-time basis from 28 November 1955. Dr. Sujit Kumar Mitra who had been away for about two years as a research assistant at the University of North Carolina, joined the Institute on 20 August 1956.

The following persons left the Institute during the year: Dr. P. B. Patnaik left in September 1955 to join the Central Statistical Organizations. Dr. G. Kallianpur was given leave to take up some work in collaboration with Professor Norbert Wiener and left for USA in May 1956. Shri Ravi Kumar left in January 1956 to join the Directorate General, Ordnance Factories as statistician. Shri Ananta Pandey left in February 1956 to join the University of Lucknow.

14. *Cost Accounts Section:* This Section was, as usual, mainly concerned with evaluating output for various operational items in terms of equivalent standard hours for all primary workers, and also for all jobs undertaken by the Projects Division. The system of incentive bonus was in operation in the form of prizes for efficient performance. The Board of Standards continued to function and held 12 meetings in the year. Reasonable levels of output rates for 238 items of work were fixed.

15. *Sankhyā: The Indian Journal of Statistics:* During the year under review four issues of *Sankhyā* were published comprising Parts 3 and 4 of volume 15 and Parts 1 and 2 of volume 16, and containing between them 21 papers including technical papers besides several important contributions on the Second Five Year Plan of India. The demand for the journal is steadily increasing and a number of new subscribers have been enrolled.

### 13. BRANCHES

#### 13.1. Bombay Branch

1. The Bombay Branch, in addition to its local activities, undertook many items of work on behalf of the parent Institute. (Names of office-bearers are given in Appendix 4).

2. *Sample Surveys:* The Branch carried out the field work of the ninth round of the National Sample Survey in Bombay City from May to November 1955. The field work for the tenth round was in progress. Analysis of data on small scale industrial establishments in the City collected last year was also in progress. The Branch published the report on the enquiry into economic conditions of middle class families in Bombay city conducted in 1950.



## TWENTYFOURTH ANNUAL REPORT : 1955-56

3. *Quality Control:* The staff of the Board of Management for Quality Control continued to work in close cooperation with the SQC Unit, Bombay. Regular visits to factories were paid by members of the Board. Apprentices of the Board helped the SQC Unit in the collection and analysis of the data from member factories.

4. *Institute Examinations:* The Statistician's Diploma Examination was conducted at the Bombay centre in August 1955 and in March 1956. Fourteen candidates appeared for the examination in August 1955 and eleven in March 1956.

5. The Branch arranged a series of lectures during the year. Among the lecturers were Shri V. P. Godambe and Prof. Oskar Lange. The Branch also received guests like Dr. Arthur Geddes, Prof. Oskar Lange, Prof. P. A. Baran, Dr. J. Links, Prof. E. A. Rowse, Mr. D. J. Desmond, Prof. J. K. Galbraith and Prof. S. S. Wilks.

### 13.2. Mysore State Branch

1. *Membership:* As on 31 March, 1956 there were 22 ordinary members, 3 sessional members and 2 life members bringing the total membership of the Branch to 27. The Second Annual General Body meeting of the Branch was held on 27 May 1955. (Names of members of the Executive Committee in Appendix 4).

2. *Statistical Surveys:* A pilot sample survey on 'the spread of labour and their mode of transport in industries in Bangalore at Hindusthan Aircraft Ltd.," was completed in March 1956. Another pilot sample survey on "consumers' demand for products of cottage industries in Mysore city" was also completed during this period. A scheme was drawn up for a survey on "job satisfaction among workers in industrial concerns in Bangalore".

3. *Visiting experts, lectures and seminars:* Among the visitors to the Branch who delivered lectures were Prof. P. C. Mahalanobis, Mr. G. Taguchi, Dr. A. Geddes, Prof. Norbert Wiener, Mr Rowse, Prof. Oskar Lange, Prof. P. Baran and Prof. H. C. Ghosh.

### 13.3. Aligarh Branch

1. The Aligarh Branch was inaugurated by Dr. Zakir Hussain, Vice-Chancellor of the Muslim University, Aligarh, on October 6, 1955 with Dr. Hussain as President, and Prof. D. P. Mukherjee of the Economics Department and Prof. S. M. Shah of Mathematics and Statistics Department as Vice-Presidents. (Names of office bearers are given in Appendix 4).

2. The Branch undertook the study of two important problems, viz., agricultural underemployment and potential development of cottage industries. A preliminary survey of village Sarsol, 2 miles away from Aligarh, was carried out with the main purpose of studying the generation of rural incomes. The summary of results was sent to the Central Statistical Organization, Government of India. A detailed survey of village Faridpur was carried out for studying different aspects of rural economy. The survey extended from 13 March to 14 April, 1956. A report was prepared on the basis of the data collected and was submitted to the Indian Statistical Institute, Calcutta.



# Indian Statistical Institute : Receipts and Payments Account

## To Receipts

	Rs.	As.	P.	Rs.	As.	P.
1. Opening Balance : Cash in hand and at Banks ..	25,096	15	1			
Unadjusted suspense of 1954-55 ..	73,902	11	11			
				98,999	11	0
2. Membership subscription .. ..				3,537	4	0
3. Training fees .. ..				4,090	0	0
4. Examination fees and other receipts .. ..				32,390	9	0
5. S.Q.C. membership and training fees including service charges .. ..				76,639	2	0
6. Block grants from the Government of India, Ministry of Finance for—						
i) Research, Training & General Purposes ..	7,26,000	0	0			
ii) International Statistical Education Centre ..	95,400	0	0			
iii) Statistical Quality Control Units ..	1,36,700	0	0			
iv) Electronic Laboratory & Computing Machines Sector ..	1,15,000	0	0			
v) Economic wing .. ..	1,00,000	0	0			
vi) Operational Research Sector .. ..	2,00,000	0	0			
vii) Multipurpose National Sample Surveys ..	38,38,700	0	0			
viii) Employment Surveys (Planning Commission's restricted programme) .. ..	17,600	0	0			
				52,29,400	0	0
7. Development grant received from Government of India, Ministry of Finance in lieu of Supervision fees ..				2,50,000	0	0
8. Funds received from Government of India, Ministry of Finance for disbursement to the trainees at ISEC selected as fellows under the Technical Cooperation Scheme, Colombo Plan .. ..				16,170	0	0
9. Arrear dues for work in earlier years from :						
i) Government of India, Ministry of Home Affairs for United Nations & Government of India joint population studies at Mysore .. ..	68,000	0	0			
ii) Government of India, Ministry of Finance for Employment Surveys integrated with NSS 9th Round .. ..	30,000	0	0			
iii) Government of India, Ministry of Finance for Planning Commission's restricted programme on Employment Surveys .. ..	1,100	0	0			
				99,100	0	0
10. Refund received through Government of India on account of excess amount paid during 1954-55 through the Indian Embassy, United States of America for preparation of a duplicate set of punched cards relating to Industrial Statistics .. ..				708	15	0
11. Receipts from other sources for small items of work :						
a) F.A.O. Rome .. ..	2,381	0	0			
b) University of Michigan .. ..	1,419	0	0			
				3,800	0	0
12. Miscellaneous receipts :						
a) Donations & contributions .. ..	5,264	0	0			
b) Sale proceeds of waste paper, cards etc. ..	2,143	1	3			
				7,407	1	3
13. Deposit accounts .. ..				5,514	5	3
14. Outstanding liabilities for goods & services ..				1,48,884	6	6
				Rs. 59,76,641	6	0

**(current expenditure) for the year ending 31st March 1956**

<i>By Payments</i>		Rs.	As.	P.	Rs.]	As.	P.
1.	Salary, dearness allowance, honorarium etc...	25,46,556	15	9	26,41,816	5	9
	Employer's contribution to workers' Provident Fund ..	95,259	6	0	1,30,183	3	6
2.	Travelling allowances .. .. .				19,826	3	0
3.	Overtime allowances .. .. .				1,45,000	0	0
4.	Contribution to leave salary fund .. .. .						
5.	Visiting Professors, Fellows, Foreign experts & scientists etc. (SQC & ORU sectors) .. .. .	1,36,445	5	9			
	Research & Training sector (Transfer to fund account) ..	45,000	0	0	1,81,445	5	9
6.	Scholarships, Stipends & assistance to trainees of the R. & T. School (transfer to fund account) .. .. .				1,00,000	0	0
7.	Non-Colombo Plan Scholarships & assistances .. .. .				14,239	12	0
8.	Disbursement of Fellowship allowance to ISEC trainees under Colombo Plan .. .. .				15,038	15	6
9.	Prizes to workers for initiative etc. .. .. .				11,600	0	0
10.	Contribution to Gratuity fund .. .. .				1,10,000	0	0
11.	Machine Tabulation Expenses :						
a)	Hire & maintenance of Tabulating equipment, Key punches and verifiers including freight, transport, etc. .. .. .	5,08,643	1	0			
	etc. .. .. .	1,80,419	2	9			
b)	Cost of cards, Cabinets etc. .. .. .						
c)	Payments to I.B.M., B.T.M., Powers-Samas & Gokhale Institute of Politics & Economics, Poona for tabulation of N.S.S. data .. .. .	1,24,816	8	6	8,13,878	12	3
12.	Printing & Publication (including paper for printing) ..				65,191	6	9
13.	Society type Activities .. .. .				25,692	13	11
14.	Examination expenses .. .. .				15,949	8	6
15.	Books & Journals (including cost of binding) .. .. .				56,224	7	11
16.	Workshop, Photo & Microfilm .. .. .				18,717	1	0
17.	Stores & materials for the Computing machine & Electronic Laboratory unit .. .. .				29,240	11	6
18.	Repairs & replacement of machineries, equipment, accessories, furniture & fittings etc. .. .. .				53,414	5	3
19.	Stationeries & consumable stores .. .. .				69,064	7	3
20.	Auditor's fees .. .. .				3,600	0	0
21.	Bank charges & interest .. .. .				8,897	8	0
22.	Crop-cutting labour charges & experiments .. .. .				3,643	3	6
23.	Telephone charges .. .. .				20,714	6	0
24.	Postage, telegram, advertisement, & other miscellaneous contingencies .. .. .				53,470	0	6
25.	Electric charges .. .. .				23,373	2	0
26.	Rent, rates & taxes—including those of field & camp offices .. .. .				1,35,438	3	0
27.	Repairs & maintenance of land & buildings including petty constructions .. .. .				97,461	9	0
28.	Transport .. .. .				48,590	3	0
29.	Workers' welfare & amenities .. .. .				81,655	0	9
30.	Development at Director's discretion .. .. .				29,648	0	6
31.	Statistical Quality Control Conference .. .. .				1,332	3	6
32.	Repayment of outstanding liabilities (as per last account) ..				63,301	5	6
33.	Repayment of Bank overdraft (as per last account) ..				65,722	3	10
34.	Repayment of loan from funded accounts (as per last account) .. .. .				1,95,000	0	0
35.	Repayment of old deposits (as per last account) .. .. .				3,963	2	3
36.	Dopreciation charges (transferred to fund a/c.) .. .. .				71,000	0	0
37.	Development Grant (transferred to fund a/c.) .. .. .				2,50,000	0	0
38.	Temporary loan to Capital Expenditure A/c. .. .. .				1,49,682	4	0
39.	Loans to staff for educational & house building purposes ..				14,984	1	0
40.	Amount under suspense with staff and others pending final adjustment .. .. .				75,010	1	2
41.	Closing Balance : Cash in hand & at Banks :						
i)	with Central Office .. .. .	36,856	15	8			
ii)	with branches & sub-offices .. .. .	26,774	5	0			
					63,631	4	8
					Rs. 59,76,641	6	0

Examined and found correct.  
Sd/- P. O. NANDI & Co.,  
Chartered Accountants & Auditors

**PART 3: APPENDICES****Appendix 1 : Members of the Council, 1955-56**

*President:* Shri Chintaman D. Deshmukh.

*Vice-Presidents:* Dr. P. N. Banerjea, Professor D. R. Gadgil, Shri K. C. Mahindra, Sir Shri Ram.

*Chairman:* Sir D. N. Mitra.

*Vice-Chairmen:* Dr. S. K. Banerji, Prof. S. N. Bose, Shri K. P. Goenka, Shri S. C. Ray.

*Treasurer:* Dr. Satya Churn Law.

*Secretary:* Professor P. C. Mahalanobis.

*Joint-Secretaries:* Shri Nihar Chandra Chakravarti and Shri S. C. Sen.

*Members:* Shrimati Chameli Bose, Prof. K. N. Chakravarti, Shri Nistaran Chakravarti, Shri V. M. Dandekar, Shri M. Ganguli, Prof. H. C. Ghosh, Shri Nimai Charan Ghosh, Dr. Q. M. Hussain, Prof. D. G. Karve, Prof. K. B. Madhava, Shrimati Nirmal Kumari Mahalanobis, Shri N. T. Mathew, Shri Mani Mukherjee, Dr. U. S. Nair, Shri Pitambar Pant, Dr. B. Ramamurti, Dr. C. R. Rao, Mr. N. Sundararama Sastry, Shri J. M. Sen, Shri Sadasiv Sengupta.

**Appendix 2 : Governing Body of the Research and Training School, 1955-56**

Shri Chintaman Deshmukh (*President, ex-officio*), Sir D. N. Mitra (*Chairman, ex-officio*), Prof. P. C. Mahalanobis (*Secretary, ex-officio*), Shri Bali Ram Bhagat and Shri C. V. Narasimhan (*Representatives of the Government of India*), Dr. N. S. R. Sastry (*Reserve Bank of India*), Dr. U. Sivaraman Nair (*Inter-University Board*), Mr. J. A. R. Tainsh (*Associated Chamber of Commerce*), Shri D. N. Mukherjee (*Federation of Indian Chambers of Commerce and Industry*), Dr. V. G. Panse (*National Institute of Sciences*), Dr. J. P. Niyogi (*Indian Economic Association*), Sir Shri Ram, Dr. S. K. Banerji, Prof. S. N. Bose, Prof. K. B. Madhava, Shri N. C. Chakravarti and Dr. C. R. Rao (*Representatives of the Council of the Indian Statistical Institute*).

*Finance Committee* (of the Governing Body): Sir D. N. Mitra (*Chairman, ex-officio*), Prof. P. C. Mahalanobis (*Secretary, ex-officio*), Dr. C. R. Rao (*Director of Research and Training School*), Shri C. V. Narasimhan and Shri S. Jayasankar (*Representatives of the Government of India*), Mr. J. A. R. Tainsh and Shri Nihar Chandra Chakravarti (*Members of the Governing Body*).

**Appendix 3: Committees set up by the Council, 1955-56**

*Finance Committee:* Sir D. N. Mitra (*Chairman*), Dr. S. C. Law (*Treasurer, ex-officio*), Prof. P. C. Mahalanobis (*Hony. Secretary, ex-officio*), Dr. S. K. Banerji, Shri S. C. Ray, Shri C. V. Narasimhan, Shri P. Pant, Shri S. C. Sen, Shri N. C. Chakravarti (*Member and Secretary*).

*Journal Committee:* Prof. S. N. Bose, Dr. Debabrata Basu, Dr. G. Kallianpur, Prof. K. B. Madhava, Prof. P. C. Mahalanobis (*Editor, Sankhyā, ex-officio*), Shri Moni Mukherjee, Dr. U. S. Nair, Shri D. B. Lahiri, Shri P. Pant, Dr. B. Ramamurthi, Dr. C. R. Rao, Dr. N. S. R. Sastry and Dr. P. B. Patnaik (*Representatives of the Governing Body*), Shri Anikendra Mahalanobis (*Member and Secretary*).



## TWENTYFOURTH ANNUAL REPORT : 1955-56

*Examinations Committee :* Dr. N. S. R. Sastry, Shri V. M. Dandekar, Prof. P. C. Mahalanobis, Shri M. L. Ganguli, Shri N. C. Ghosh, Shri P. Pant, Dr. B. Ramamurti, Shri N. C. Chakravarti, Dr. C. R. Rao, Shri J. M. Sengupta, Shri Sadasiv Sengupta, Shri D. Y. Lele, Prof. K. B. Madhava, Dr. S. K. Banerji, Dr. U. S. Nair (*Representative of the Governing Body*) and Shri J. M. Sen (*Member and Secretary*).

### Appendix 4 : Office Bearers and Council Members of Branches, 1955-56

*Bombay :* Shri V. L. Mehta (*President*), Prof. C. N. Vakil, Shri R. G. Saraiya, Shri L. S. Vaidyanathan and Dr. N. S. R. Sastry (*Vice-Presidents*), Dr. K. S. Rao and Shri K. C. Cheriyan (*Joint Secretaries*); Dr. D. T. Lakdawalla (*Treasurer*), Dr. R. L. N. Iyenger, Prof. M. C. Chakravarti, Shri H. T. Parekh, Shri A. S. Palekar and Shri M. A. Telang (*Members of the Council*).

*Mysore :* Prof. S. K. Ekambaram (*President*), Shri A. Ananthapadmanabha Rao (*Vice-President*), Shri Srinagabhusana (*Secretary*), Sri R. Guraraja Rao (*Joint Secretary*), Shri R. Ramaswami (*Treasurer*), Shri Ravi L. Kirloskar, Shri S. K. Rama, Shri H. S. Narayana Rao, Shri M. C. Satyanarayana, Shri R. Natarajan and Shri M. V. Venkataraman (*Members of the Council*).

*Aligarh :* Dr. Zakir Hussain (*President*), Prof. D. P. Mukherji (*Vice-President*), Prof. S. M. Shah (*Secretary*), Shri Harish Ch. Gupta, Mr. Abu Salim, Shri A. R. Kokan, Shri S. S. Gupta, Shri. K. A. Naqvi, Shri M. A. Raj, Shri Abdul Qayum, Shri M. A. Rizvi, Shri M. Sultan and Mrs. Syera Irfan (*Members of the Council*).

### Appendix 5 : Scientific Enquiries

#### Sponsors and Subjects

MISS A. DASGUPTA, *Government Training College for Women, Simla :*

Correlation between Mental age and scores in arithmetic of Secondary School students.

DEPARTMENT OF APPLIED PSYCHOLOGY, *University of Calcutta :*

Differential effect (in respect of age, sex and levels of intelligence) of practice on the score in a psychological test.

Differential performance of delinquent and normal children in psychological tests.

Effect of time interval on recall in the case of nonsense and meaningful syllables.

Reminiscence effect in whole and part learning of paired associates.

DEPARTMENT OF PSYCHOLOGY, *University of Gauhati :*

The influence of colour in the assessment of lengths.

AGRICULTURAL COLLEGE, *Government of West Bengal :*

Analysis of manurial and varietal experiments on paddy.

NATIONAL MEDICAL COLLEGE, *Calcutta :*

Statistical analysis of the effect of a new drug on Cholera.

RICE RESEARCH INSTITUTE, *Cuttack :*

Estimation of linkage between certain factors in paddy.

CITY COLLEGE, *Calcutta :*

Land utilization statistics for the different states of India;

CALCUTTA PURE DRUG COMPANY, 2, Cooper Lane, *Calcutta :*

Certain vital statistics relating to India.

### Appendix 6 : Sampling Design of the 9th and 10th rounds of the NSS

1. *Ninth Round* : In the 8th round there were 1242 sample villages (central sample only) which increased in the 9th round to 1624, or by about 14 per cent. The state samples were concerned with land holding surveys etc., in the 8th round only, and are therefore excluded from the comparison with the size of the sample in the 9th round. In the urban sector in the 8th round 444 urban blocks (central sample) were selected, and this number was increased in the 9th round by 2108 or nearly five times.

2. The sample villages were allocated to districts or groups of districts, which were the ultimate strata, in proportion to their relative rural population. The allocated numbers were so adjusted as to make them multiples of four. Samples were then drawn from these strata at random with probability proportional to population and with replacement. In the urban sector the sample blocks, which were the first stage units, were allocated to States, in the first instance, in proportion to their respective non-agricultural population. Within each State, the State quota was further allocated to each of the big cities with population (1951 census) of 3 lakhs and above and capitals of part A and part B States with population below 3 lakhs, except in the case of Shillong, capital of Assam, and to the rest of the urban area in that State. The ultimate strata were the individual cities mentioned above, and the remaining urban area within a natural division of a State. Some deviations from the above rule of stratification had however to be made with respect to the area known as "Greater Calcutta", exclusive of the cities of Calcutta and Howrah. This area was kept separate from the remaining urban areas of West Bengal. Similar was the case with "Greater Bombay", less the city of Bombay. The allocation of the State quota to the strata so formed was made proportional to the respective non-agricultural population (1951 Census), and preferential weights in varying degrees were assigned to the strata constituted by individual cities. Adjustments were further made to make the strata allocations multiples of 4 in all cases. Within each stratum the required number of sample blocks were drawn according to the method of systematic selection taking a random start and completing the cycle. Four such systematic samples of blocks were taken so as to provide four independent sub-sample estimates.

3. The State sample of Bombay consisted of as many sample villages and blocks as the central sample in the State, excepting for Bombay City proper, where the size of the State sample was half the size of the Central sample. The extended State sample of U.P. numbered only half the total of sample villages and blocks constituting the Central sample in the State. As regards the subject coverage, Bombay covered all the schedules as in the Central sample, but U.P. took up two enquiries only, namely employment and unemployment, and household manufacture and handicrafts.

4. *Tenth Round* : The same set of 1624 sample villages selected for the 9th round was surveyed for all subjects of enquiry in this round. In addition, a sample of 3260 villages which was divided into 5 independent sub-samples, was selected for land utilization survey and of these 5 sub-samples, villages of sub-samples 3, 4 and 5 were selected for crop-cutting experiments. In the rural sector, out of 2108 sample blocks selected for the 9th round, 1328 blocks were surveyed in the 10th round.

5. For the 1614 villages and 1328 urban blocks the design remained the same as in the previous round. For the 3260 villages the districts usually formed the strata.



## TWENTYFOURTH ANNUAL REPORT : 1955-56

Allocation to different strata was made by a joint consideration of the geographical area of the stratum and the proportion of the area under various crops for the season covered by the present round. Selection of villages within each stratum was done with probability proportional to area and with replacement. In cases where area figures were not available, selection had to be made at random with equal probability.

### Appendix 7 : Machine Tabulation Unit

#### 1. The strength of Tabulation Equipments at different centres

period	machines	Baranagar		Giridih	Delhi		total
		IBM/Holl.	Power-samas		IBM/Holl.	Powers-samas	
Apri 1 to Dec. '55	Accounting machine	8	2	1	—	—	11
	ESM (101)	2	—	—	—	—	2
	Sorter	10	2	1	—	—	13
	Multiplier	2	1	—	—	—	3
	Collator	2	—	—	—	—	2
	Reproducer	6	—	—	—	—	6
Jany. to Mar. '56	Gang Punch	3	2	1	—	—	6
	Accounting machine	9	2	1	1	1	14
	ESM (101)	2	—	—	—	—	2
	Sorter	13	2	1	1	2	19
	Multiplier	2	1	—	1	—	4
	Collator	2	1	1	1	—	5
Mar. '56	Reproducer	7	—	1	—	—	8
	Gang Punch	5	2	1	1	—	9

2. *The output:* The per hour output from tabulating units recorded an increase from 0.95 to 1.39 for accounting machines and from 4.62 to 5.43 for ESMs. The total card-passage through all the tabulating units and ESMs during the year was 16,327 thousands and 11,308 thousands respectively. In spite of the improved output record for the year, it was felt that further improvement in output record can be achieved by effecting some changes in the composition of punched card tabulating equipments by giving optimum number of auxiliary machines for tabulating units. A sub-committee was formed in February 1956 to effect necessary changes.

3. Unlike previous years, card-stores were located at Baranagar, Giridih and Delhi with 250 lakhs of cards stored in cabinets with 50 drawers and 25 drawers. This constitutes the punched cards of 2nd to 9th rounds of NSS and related surveys. Besides, about 5 lakhs of Y-sample cards of 1941 census stored at Giridih were brought down to Baranagar for MIT work.

4. During the year under review, method of scoring by the PRS Unit conducting individual tests of aptitude and ability was mechanized through the probability method of sorting by electronic statistical machines. This has widened the field of application of ESMs.



**Appendix 8 : Library**

1. The Central Library containing about 56,000 books and monographs and 1190 journals besides a large collection of special material was located at Baranagar. Service centres were maintained at the City Office at 9/B Esplanade East, Calcutta, and at Giridih. Transfer of material and personnel from 204 Barrackpore Trunk Road, where the library was partially accommodated was taken up with a view to concentrating it in the main premises at 203 Barrackpore Trunk Road as well as to make room for the proposed Museum at the former place.

2. *Books* : The library acquired 2717 volumes of books against 1945 last year. Of these 552 were received as gifts, 53 in exchange and 82 through Review Section. In addition to these, 719 volumes from the Late J. C. Sinha's collection which was purchased last year were integrated into the library stock.

3. *Periodicals* : The library received 1190 periodicals and annuals against 1126 last year. Of these 257 periodicals were subscribed, 354 were received as gifts and 579 were received on exchange basis. The library subscribed 42 new journals and entered into exchange arrangements with 4 new Indian and 8 foreign agencies among which 1 was in East Germany, 1 in the USA, 1 in Spain, 1 in Switzerland, 1 in Pakistan and 3 in France.

4. *Special Gifts* : The library thankfully accepted a concession of \$ 125 offered by the Joint Committee on Slavic Studies (American Council of Learned Societies and the Social Science Research Council) towards the annual subscription of \$ 150 for the Current Digest of the Soviet Press.

The United States Information Services offered one year's subscription to the New York Times (International Edition) Sunday Issues only, which was thankfully accepted.

5. *Bibliographical Services* : Two bibliographies were compiled—one on Cottage Industries and the other on Industrial Management and Industrial Technology.

The library continued to issue the weekly list of selected periodicals and the monthly bulletin of new acquisition. The issue of the Index to Current Periodicals had to be kept suspended for want of adequate technical staff.

6. *Service and Circulation* : The number of library members increased to 758 from 624 last year. The total number of books, journals and other materials issued was 38,055 against 23,033 last year, of which 28,211 were issued from the Reference Section and 9844 from the Lending Section. The total number of requests received was 43,120, so that nearly 11.5 per cent of the requests could not be fulfilled, against 6 per cent last year. The rise in the number of unfulfilled requests was mainly due to increased demand for text books from the students of several new training courses.

7. *Circulating Library* : There was new acquisition of 733 Bengali, 152 English, 26 Hindi and 38 Oriya books bringing the total to 6581 volumes. Stocks were regularly rotated amongst the branches of the library. The number of books issued from Baranagar, Calcutta and Giridih were 14,993, 1240 and 3384 respectively.

## TWENTYFOURTH ANNUAL REPORT : 1955-56

8. *News Clippings* : With the formation of the Planning Division and the publication of the Draft Plan-frame growing need was felt for easy and quick reference to topics relating to statistics and planning in India which appeared in important dailies and periodicals. For this purpose a new unit was started in June 1955 for systematic processing and indexing of relevant material.

Clippings were gathered from 5 dailies of Calcutta and the cuttings received through the All India Press Cutting Service which covered more than 100 dailies and periodicals all over India. The total number of clippings processed and properly indexed was 11,559. The number of articles indexed from periodicals was 1625.

9. *Translation* : Several requests were received and complied with for English translation of various matters—letters, articles, books in other foreign languages, specially French and Russian. A major work taken up was the translation of three Russian books on the introduction of metric system in the USSR.

10. *Records Unit* : The Records Section continued to function at 206 Barrackpore Trunk Road, Calcutta. The Project Unit which maintains schedules, working papers and reports relating to surveys and projects carried out in the Institute arranged, classified and indexed 5004 files, against 4387 last year, bringing the total number of files thus arranged so far to 12,429.

In the map unit, 84,249 cadastral survey maps, 251 P.S. maps and 10 district maps of West Bengal as well as 6185 State maps of Indian Union, totalling 90,668 sheets, have been finally processed and shelved through serial sorting and numbering. The number of maps issued to Field Branch and other units was 1866.

11. *Photographic Unit* : Major work on documentary reproduction consisted of 2588 frames of microfilms, 1359 paper prints from microfilm and 1038 photostats. The Photographic Unit also took 1489 still photographs of individuals, groups and important functions at the Institute as well as 1200 feet of motion picture. In addition to these, 2766 bromide enlargements were made.

Other types of work included preparation of profiloscope pictures, acre plates, projection slides (black and white, and colour) etc., involving about 220 exposures. Requests for documentary reproduction from other Institutions were also complied with by supplying microfilms, photostats and photoprints.

To supplement the existing equipment the library acquired a Remington Rand Transcopy unit for preparing multiple copies of documents and material within a short time.



**Appendix 9 : List of papers completed during 1955-56****A. Theoretical Statistics****A.1 SUMMARY OF RESEARCH WORK IN THEORETICAL STATISTICS**

(The figures within brackets refer to the list of papers appearing later in this Appendix.)

1. *Tests of significance (3,21,22,23,24,33)*: Considering the conditional probability given the total of a number of observations from a Poisson distribution some exact tests of significance were constructed for judging (i) goodness of fit, (ii) homogeneity of observations, (iii) deviations in the frequency of the 'zero' class. Tables have been provided in a limited number of cases. The same technique was exploited in obtaining some exact tests of significance in the case of a binomial distribution also.

The concepts of partial and multiple correlations associated with a multi-normal population have been extended to the case of populations admitting a multiple classification as in a contingency table and large sample tests for hypotheses concerning them have been developed. Analogous problems of testing 'main effects and inter-actions' in cases of sampling from multinomial populations have been discussed. Limiting power functions (in Pitman's sense) of the frequency chi-square tests have been obtained and their possible uses indicated.

2. *Design of experiments (2,11,12,13,20,27,28,29,30,35,36,37)*: 'Orthogonal arrays' have been used to construct fractional replicated designs for asymmetrical factorial experiments. A new class of arrangements called partially balanced arrays have been introduced and their use in industrial experimentation demonstrated.

The conditions on the design matrix for certain parametric functions to uncorrelated estimates have been investigated.

With reference to varietal trials a number of contributions were made both on the construction of designs and analysis. A new system of arrangements known as the Quasi-block designs, which occur as a sub-class of PBIB designs has been introduced. All linked block designs with number of replications and block size less than 10 have been enumerated. A complete enumeration of all two associated PBIB designs involving three replications have been carried out. By considering the dual of group divisible designs with block size two, a very useful class of PBIB designs with two replications and a maximum of 5 different errors has been obtained. On the analysis side simple methods of carrying out inter-block analysis have been suggested. It was shown that the expressions for inter-block analysis can be obtained from the corresponding expressions of the inter-block analysis by just changing the parameters of the designs in a simple way. This is applicable in whatever way intra-block analysis is carried out by estimating the treatment contrasts first or otherwise by estimating the block contrasts first.

3. *Estimation (8,26,32)*: The efficiency of estimating parameters by the method of moments has been investigated in the case of Poisson and Binomial distributions, truncated at zero. The loss of efficiency is not serious while the method of estimation is simple compared to the labour involved in obtaining maximum likelihood estimates. Extensive tables have been provided to obtain likelihood estimates in a simple way.

Under conditions of Fisher consistency and Frechet differentiability lower limit to the asymptotic variance of a statistical functional is shown to be the information limit, information being as defined by Fisher.



## TWENTYFOURTH ANNUAL REPORT : 1955-56

4. *Sample Surveys (34,38)* : The variance of the estimate of the population total in a multi-stage survey where units are chosen with probabilities proportional to size is decomposed into meaningful stage components and their unbiased estimates are derived.

A simple sampling scheme has been devised by which at most  $(n-1)$  sample units are examined for two sampling enquiries involving two systems of p.p.s. sampling for the units, which ordinarily require the examination of  $2n$  sampling units. A great reduction in cost can be achieved with no loss in precision.

5. *Characterization theorems (1,9,10)* : Work on characterization of distribution functions mentioned in the earlier reports was continued. An important characterization of the normal distribution was obtained using the properties of distributions of linear statistics.

Using the necessary and sufficient condition for a  $k$ -dimensional vector to be distributed as multivariate normal, that every linear function should be univariate normal, several characterization theorems of the multivariate normal distribution have been obtained by first reducing the problem to the univariate case for which a solution is readily available.

6. *Stochastic processes (4,5,6,15,16)* : Making use of the canonical representation of an infinitely divisible distribution due to Kolmogorof and Polya it was proved that an infinitely divisible distribution is necessarily bounded.

Some theorems on the limiting distributions of maximum of partial sums have been proved assuming the existence of the second moment. This generalizes all the results of Chung obtained on the assumption of the existence of the third moment.

7. *Miscellaneous (7,14,25,31)* : There were a number of other contributions relating to the construction of analogue machines for solving linear equations, mathematical models of planning, acceptance sampling for variables and problems of optimum selection in multivariate analysis.

### A.2. PAPERS

1. BASU, DEBABRATA : A note on the multivariate extension of some theorems related to the univariate normal distribution. (submitted to *Sankhyā*).
2. CHAKRAVARTI, INDRAMOHAN : Fractional replication in asymmetrical factorial designs and partially balanced arrays. (*Sankhyā*, 17, 143).
3. ——— and C. R. Rao : Some small sample tests of significance for a Poisson distribution. (*Biometrics Bull.*, Sept. 1956).
4. CHATTERJEE, SRISTIDHAR : A note on mean first passage and recurrence times.
5. ——— : A generalization of a result due to Chung.
6. ——— and R. P. Pakshirajan : On the unboundedness of the i.d. law. (submitted to *Sankhyā*).
7. DES RAJ : On optimum selections from multivariate populations. (*Sankhyā*, 14, 363).
8. KALLIANPUR, GOPINATH and C. R. Rao : On Fisher's lower bound to asymptotic variance. (*Sankhyā*).
9. LAHA, RADHA GOVIND : On a characterization of the normal distribution from properties of suitable linear statistics. (*Ann. Math. Stat.*, in press).

10. ——— : Characterization of probability distributions and statistics. (Thesis submitted to Calcutta University for the degree of D. Phil.).
11. ——— and J. Roy : Classification and analysis of linked block designs. (*Sankhyā*, 17, 115).
12. ——— and J. Roy : Partially balanced linked block designs. (*Sankhyā*, in press).
13. ——— and J. Roy : Two associated PBIB designs involving three replications. (*Sankhyā*, 17, 175).
14. MATTHAI, ABRAHAM : A class of acceptance sampling plan for variables. (*Bull. of the Quality Control Association*, Bangalore, Vol. III).
15. PAKSHIRAJAN, RAJAKULARAMAN PONNUSAMI : On the maximum partial sums of sequences of independent random variables.
16. ——— and S. D. Chatterjee : On the unboundedness of the i.d. law. (Submitted to *Sankhyā*).
17. MAHALANOBIS, P. C. : The approach of operational research to planning in India. (*Sankhyā*, 16, parts 1 & 2).
18. ——— : Approach to planning in India. (Based on a talk broadcast from All India Radio on 11 September, 1955).
19. ——— : Statistics must have purpose. (Presidential address to the Third Pakistan Statistical Conference, Lahore, February 1956).
20. MITRA, SUJIT KUMAR : A note on orthogonality and design of experiments. (*Sankhyā*, in press).
21. ——— : On Bartlett's test of complex contingency table interaction. (*Sankhyā*, in press).
22. ——— : On the limiting power function of the frequency Chi-Square test. (*Sankhyā*, in press).
23. ——— : Contributions to the statistical analysis of categorical data. (North Carolina Institute of Statistics Mimeograph Series No. 142, December 1955).
24. ——— and S. N. Roy : An introduction to some non-parametric generalizations of analysis of variance and multivariate analysis. (*Biometrika*, in press).
25. MITRA, SAMAR : Electrical analogue computing machine for solving linear equations and related problems. (*Review of Scientific Instruments*, 1, 26, 453, May 1955).
26. PATIL, G. P. : The efficiency of the "Two-Moments" estimate of the parameter in a single truncated binomial distribution.
27. RAMAKRISHNAN, C. S. : The dual of a two associate PBIB design and new designs with two replications. (*Sankhyā*, 17, 133).
28. ——— : Designs with two replications as duals of certain group divisible designs.
29. RAO, C. RADHAKRISHNA : On the recovery of inter-block information in varietal trials. (*Sankhyā*, 17, 105).
30. ——— : A general class of Quasi-factorial and related designs. (*Sankhyā*, 17, 165).
31. ——— : Analysis of dispersion with missing observations. (*Jour. Roy. Stat. Soc.* Sept., 1956).

32. ——— and G. Kallianpur : On Fisher's lower bound to asymptotic variance. (*Sankhyā*).
33. ——— and I. M. Chakravarti : Some small sample tests of significance for a Poisson distribution. (*Biometrics Bull.*, Sept., 1956).
34. ROY, JOGABRATA : Variance components in multistage PPS sampling. (*Sankhyā*, in press).
35. ——— and R. G. Laha : Classification and analysis of linked block designs. (*Sankhyā*, 17, 115).
36. ——— and R. G. Laha : Partially balanced linked block designs. (*Sankhyā*, in press).
37. ——— and R. G. Laha : Two associates PBIB designs involving three replications. (*Sankhyā* 17, 175).
38. ROY CHOUDHURY, D. K. : An integration of several PPS surveys. (*Science & Culture*, Vol. 22, No. 2).

## B. Applied Statistics

### B.1. BIOMETRIC STUDIES

39. DAS, BHUPENDRA CHANDRA and Nalvandov, A. V.: Responses of prepuberal chicken ovaries to avian and mammalian gonodotrophins. (*Endocrinology*, 57, 705).
40. ROY, SUBODH KUMAR : Studies on the activities of earthworms.
41. VERMA, V. K. : A note on the relative length of fingers in a group of prisoners in a U.P. Jail.
42. ——— : Observations taken on the ear lobes of a group of U. P. convicts.
43. ——— : A note on the human face-observations on convicts.

### B.2. PSYCHOMETRIC STUDIES

44. CHATTERJI, S. : Machine scoring of objective tests.
45. DAS, RHEA S. : A logical analysis of the concepts 'personality' and 'attitude'. (Submitted to *Psychological Review*).
46. ——— : Recommendations for personnel selection in India based on British selection methods in Civil Service & in Industry. (Submitted to *Sankhyā*).
47. ——— and J. P. Das and R. Rath : Understanding versus suggestion in the judgement of literary passages. (*Journal of Abnormal and Social Psychology*, 1955).
48. DAS GUPTA, B. : Some aspects of scaling in the two-stage selection process.
49. ——— : A simplified method of item analysis.
50. ——— and others : The validity of the Vellore Medical College selection methods. (*Sankhyā*).
51. HARPER, A. EDWIN, Jr. : Modern objective examination marking. (Cyclostyled).
52. ——— : A manual of test scoring and edge marking for item analysis for use with the Harper-Mitra Answer Sheet.
53. ——— and others : The validity of the Vellore Medical College selection methods. (*Sankhyā*).
54. SUSHIL KUMAR : Development of a new edge-marking method for psychological uses.



55. MITRA, SHIB K. : Roles available to psychologists in the draft plan-frame of the Second Five Year Plan. (*Psychology and Education*, Baroda, 1956).
56. ——— : On correction for chance success.
57. ——— and D. W. Fiske of University of Chicago and J. Osterweil of the Meninger Foundation, Topeka, U.S.A. : The relationship between variability and group frequency of responses.
58. SANGAL, S. P. and others : The validity of the Vellore Medical College students selection methods. (*Sankhyā*).
59. ——— : Some aspects of problem of validity study in selection programmes.
60. SEN, TAPAS KUMAR : A systematic method of calculating discriminant functions for psychological data.

C. Working Papers on Economic Planning : 1955-56

*Working Paper Series :*

54. BARAN, P. A. : The concept of the economic surplus.
50. BETTELHEIM, CH. : Foreign trade and planning for national economic development.
52. ——— : Planned economic growth and foreign trade.
43. DIVISIA, F. : Technical coefficients (consumption units physical), France 1949.
53. GOPALAKRISHNAN, K. P. : Financing the Second Five-year Plan.
37. LANGE, O. : Some problems concerning economic planning in under-developed countries.
38. ——— : Fundamentals of economic planning.
56. ——— : Some observations on input-output analysis.
44. LOBEL, E. and DAS, P. : Productive capacity of large-scale industries in India.
46. MUKHERJEE, M. : Value of material production in India (1948-49).
47. ——— : On the shift of employment from capital intensive production processes.
45. MUKHERJEE, M. and DUTTA, UMA : A note on the ratio of increment of national income to investment.
49. ——— : An attempt at estimating parameters of a simple model of economic growth.
39. MOSKVIN, P. M. : Basic problems of the statistics of the national income in the USSR.
41. ——— : Balance of the national economy.
55. NARSU, A. V. : Introduction to a study of industrial, non-industrial and financial groupings in India (Calcutta region).
36. PISAREV, I. Y. : Balance method in the Soviet socio-economic statistics.
40. ——— : Statistics and planning.
42. ——— : The most important categories, concepts, definitions of Soviet State Statistics of Population and Industry based on 'Dictionary-Reference Book of Social Economic Statistics' of the Central Statistical Department.
57. ROY, A. and BANERJEE, N. : Monopolies and concentration of economic power—Part I.
51. RUDRA, A. : A scheme of calculations for an Annual Plan.
48. SENGUPTA, S. S. and BOSE, D. K. : A study in maximalisation of employment.

D. Miscellaneous Observations on Indian Planning

- BARAN, P. A : Reflections on planning of the economic development of India.  
 BETTELHEIM, CH. : Remarks on the second Five Year Plan : Draft outline.  
 GALBRAITH, J. KENNETH : Economic planning in India, Five comments.  
 LANGE, O. : Observations on the Second Five Year Plan. Long-term and short-term capital.  
 ROY, A. and BANERJEE, N. : The position of foreign capital in big joint-stock companies incorporated in India.  
 RUDRA, A. : Food production targets in the Second Five Year Plan.  
 SENGUPTA, S. S. : Macro economic dynamic programming.  
 STRACHEY, J. : A note on Indian development.  
 TALUQDAR, S. N. : Planning and integration in coal.  
 TINBERGEN, J. : Accounting interest rates and accounting wage rates.  
 ——— : A note on employment policy.  
 ——— : On the optimum use of the factors of production.  
 ——— : How to split up a plan into its geographical components.  
 ——— : The socialistic pattern of society.  
 ——— : The need for a uniform method of appraising investment projects.  
 ——— : A possible set up for a planning model with supply and demand equations.  
 ——— : A note on employment policy.  
 ——— : The optimum rate of saving.

**Appendix 0 : List of Trainees**

(a) Three-year (formerly two-year) Statistician's Course

(i) *Trainees attending first year class on 1st April 1955.*

1. Rattan Chand Arora; 2. Satnam Das; 3. Chandra Shekhar Dutt; 4. P. Gopalakrishnan; 5. Govind Ram Gupta; 6. Jai Prakash Gupta; 7. Satish Mohan Kansal; 8. Kaushalendra Kumar; 9. Gyanendra Deo Misra; 10. Amar Nath Nankana; 11. Vinod Prakash; 12. Miss G. Premlata; 13. Keshav Roop Rai; 14. G. Divakara Rao; 15. V. Subramaniaswamy; 16. Tilakraj Talwar.

(ii) *Trainees attending Second Year Class on 1st April 1955.*

1. S. Ramanatha Iyer; 2. G. K. Nair; 3. M. Narasimhamurti; 4. B. V. Ramasarma; 5. N. Sen; 6. V. K. Sethi; 7. T. N. Srinivasan; 8. Y. P. Bhasin; 9. G. Hariharan; 10. R. C. Jain; 11. Miss R. Kastoori; 12. O. P. Kukreja; 13. Miss N. S. R. Nanjamma; 14. S. Rajagopal; 15. S. P. Sangal; 16. H. C. Sharma.

[All the above 16 students passed the course in June 1955]

(iii) *Trainees admitted to the First Year Class in July 1955.*

1. Acharyya, S. C.; 2. R. K. Agarwal; 3. P. P. Arya; 4. L. L. Assudani; 5. S. Balakrishna; 6. K. Bhanumurty; 7. Y. K. Bhat; 8. P. K. Bhattacharyya; 9. S. K. D. Chaudhury; 10. K. M. Das; 11. D. Deverajan; 12. V. B. Dexit; 13. J. C. Gupta; 14. M. Dutta; 15. D. R. Handa; 16. Iyengar, N. S.; 17. P. K. Jain; 18. S. Khosla; 19. H. Krishnamurty; 20. D. M. Mahamunulu; 21. S. P. Malik; 22. V. Muralimohan; 23. J. L. Janda; 24. V. P. Narula; 25. P. Nath; 26. J. Prakash; 27. K. Raghavan; 28. M. L. Raina; 29. M. Ramachandran; 30. S. Ramamorthy; 31. R. Ranga Rao; 32. G. N. Rao; 33. M. L. N. Rao; 34. P. B. Rao; 35. M. G. Sardana; 36. S. N. Sawhney; 37. K. Singh; 38. A. Swarup; 39. Teekarao; 40. S. Vasudevan; 41. M. Venkatachari; 42. P. D. Verma; 43. Viswanath, R.; 44. L. A. Chandrasekharan; 45. D. K. Datta Majumdar; 46. A. S. Pankajakshan; 47. V. V. Rao.

(iv) *Trainees promoted or directly admitted to the Second Year Class in July 1955.*

1. Rattan Chand Arora; 2. Satnam Das; 3. P. Gopalakrishnan; 4. Govind Ram Gupta; 5. Jai Prakash Gupta; 6. Satish Mohan Kansal; 7. Kausalendra Kumar; 8. Amar Nath Nankana; 9. Vinod Prakash; 10. G. Divakara Rao; 11. V. Subramaniaswamy; 12. \*K. Sundararajan; 13. \*Sayed Hamid Pasha; 14. \*M. Ramakrishna; 15. \*K. Gopalakrishnamurti; 16. \*Saurindra Kumar Chakravorti; 17. \*Amar Sundar Roy; 18. \*Ramaswami Subramanian Ganesan; 19. K. P. Geethakrishnan; 20. S. Somasundar.

[\*Directly admitted to the Second Year Class].

## (b) Short-term Statistician's Course

(1) *First Session: December 1955—March 1956:*

1. N. S. Raja Rao; 2. A. K. Rai Choudhury; 3. Jamini Kanta Ghosh; 4. Nishith Ranjan Chaudhury; 5. \*Nirmal Kanti Dasgupta; 6. \*Srish Chandra Basu Roy; 7. \*Nirmalendu Bhowmik; 8. \*Kalyan Kumar Dasgupta; 9. \*Prakash Chandra Kundu; 10. \*Amalendu Sengupta; 11. Brij Bhusan Pande; 12. Kamal Kumar Pradhan; 13. K. Gobindan Kutty; 14. Biswanath Chakravorti; 15. Asoke Kumar Dutta; 16. Manik Chandra Ganguly; 17. K. Sripathi Rao; 18. Makhan Chandra Bhattacharyya; 19. Sudhir Kumar Basu; 20. Santosh Kumar Seal; 21. \*Jnanatosh Chatterjee; 22. Subhendra Kumar Banerjee; 23. \*Santimoy Banerjee; 24. Sisir Ranjan Sengupta; 25. Prabir Kumar Sandell; 26. Bhupendra Nath Bhatia; 27. Anadinath Mukherjee; 28. Amiya Kumar Bagchi; 29. \*Ram Prakash Sharma; 30. Mahinder Singh; 31. Kalyan Kumar Gupta; 32. Miss Shanti Dutta; 33. \*Manilal Ganguly; 34. Tilakraj Talwar; 35. Ajoy Kumar Gupta; 36. Miss Ela Romola Mukherjee; 37. Miss Dipa Lahiri; 38. Miss Attivilli Sita Devi; 39. Miss Padma Dutta Roy; 40. Ishan Kumar Chatterjee; 41. \*M. S. Katchapenswaran; 42. Dipak Sanyal; 43. Manajit Banerjee; 44. <sup>1</sup>Mrs. Bina Roy; 45. Shyam Chandra Sarkar; 46. Jiban Krishna Ghose; 47. <sup>1</sup>T. D. Srinivasan; 48. <sup>1</sup>D. S. Murty; 49. Abinash Chatterjee; 50. Sukhendu Nath De; 51. Sahrudin; 52. Kuldeep Singh; 53. Miss Parbati Chatterjee; 54. Debaprasad Mukherjee; 55. <sup>1</sup>George Issac.

(2) *Second Session: March—September 1956:*

1. Pratul Kumar Bagchi; 2. Viswanath Nikore; 3. Ajit Narayan Bose; 4. Valjee Jivandas Suraiya; 5. N. S. Raja Rao; 6. Niranjana Sil; 7. Srish Chandra Roy; 8. Qimat Rai; 9.

\* Certificates were awarded for successful completion,

<sup>1</sup> Auditors.



## TWENTYFOURTH ANNUAL REPORT : 1955-56

Kamal Kumar Basu; 10. Santosh Kumar Seal; 11. Harihar Mittra; 12. Jyoti Prasanna Roy; 13. T. G. Ramasubramanian; 14. G. P. Mukherjee; 15. Shankar Bhaduri; 16. Bhupendra Nath Bhatia; 17. Asoke Kumar Dasgupta; 18. E. K. Sukumaran Nair; 19. Miss Ela Romola Mukherjee; 20. Sanjoy Kumar Lahiri; 21. Ranajit Kanta Lahiri; 22. Prasaun Sengupta; 23. Bharat Kumar Kar; 24. Sukhendu Nath De; 25. Kultar Singh Bedi; 26. Dharam Vir Gulati; 27. Santosh Kumar Dutta; 28. Gopal Chandra Biswas; 29. Sujit Kumar Sarkar; 30. Sat Paul Bhatia; 31. Annabattula Jayaram; 32. Miss Shanti Dutta.

### (c) Computer's Training Course

#### (1) Session: July—December 1955:

*Junior (Morning):* 1. Sanat Kumar Das; 2. Samarendra Barua; 3. Amiya Kumar Sinha; 4. Kamal Kumar Bose; 5. Parimal Mukherjee; 6. Bhim Chandra Mitra; 7. Brojagopal Banerjee; 8. Paramesh Kumar Das; 9. Manindra Narayan Choudhury; 10. Santipriya Bhowmick; 11. Mukendeswar Bhattacharjee.

*Junior (Evening):* 1. Asoke Kumar Dasgupta; 2. Sanat Kumar Bose; 3. Ranjan Kumar Bhattacharyya; 4. Ajit Kumar Chatterjee; 5. Timir Prakash Deb; 6. Nirmal Krishna Choudhury; 7. Pronob Kumar Mitra; 8. Sunil Kumar Deb; 9. Sukha Ranjan Bhattacharjee; 10. Guru Narayan Samanta; 11. Dhiraj Lal Roy Chowdhury; 12. P. Madusudan Babu; 13. Nani Gopal Nag; 14. Golakendu Ghosh; 15. Santi Bhuson Roy; 16. Sanath Chatterjee; 17. Ajit Kumar Roy Choudhury; 18. Bibash Chandra Banerjee; 19. Sarit Kumar Raha.

*Senior (Evening):* 1. Subhas Chandra Roy Choudhury; 2. Ram Chandra Mitra; 3. Bhabatosh Sen; 4. Sanjit Banerjee; 5. Shankar Bhaduri; 6. Pranabananda Bhaduri; 7. Asoke Kumar Das Gupta; 8. Rajani Kumar Bhattacharjee; 9. Paramesh Kumar Das; 10. Dulal Chandra Dey; 11. Ramkrishna Chakravorti; 12. Nihar Ranjan Saha; 13. Braja Gopal Banerjee; 14. Nirranjan Dhar; 15. Subimal Deb; 16. Sunil Kumar Chatterjee; 17. Paritosh Mitra.

#### (2) Session : January—June 1956:

*Junior (Morning):* 1. Amarendra Nath Ghosh; 2. Animesh Chakravorti; 3. Hirendra Lal Kar; 4. Dulal Kanti Choudhury; 5. Ananta Kumar Roy; 6. Anup Kumar Sen; 7. Chittaranjan Chatterjee; 8. Subhas Chandra Raichoudhuri; 9. Sunil Kumar Bhattacharjee; 10. Ram Chandra Mitra; 11. Manindra Chandra Basu; 12. Tushar Kanti Banerjee; 13. Shyamal Chandra Roy; 14. Amal Kanti Bhattacharyee; 15. Nirmal Chakravorti.

*Junior (Evening):* 1. Baren Kumar Bose; 2. Tushar Kanti Banerjee; 3. Amrita Lal Halder; 4. Kamal Behari Paul Choudhuri; 5. Dharendra Chandra Paul; 6. Samar Krishna Basu; 7. Jyotirmoy Basak; 8. Brijesh Mohan; 9. Amal Kumar Banerjee; 10. Sailendra Narayan Gupta; 11. Saumyendra Nath Roy; 12. Madhab Chandra Munshi; 13. Raj Kumar Nandi; 14. Sanat Kumar Banerjee; 15. K. Ramalingam.

*Senior (Evening):* 1. Barun Kumar Bose; 2. Amiya Kumar Sinha; 3. Tushar Kanti Banerjee; 4. Sanat Kumar Das; 5. Samar Krishna Basu; 6. Dulal Kanti Choudhuri; 7. Sanat Kumar Bose; 8. Brijish Mohan; 9. Sunil Kumar Deb; 10. Dhiraj Lal Raichoudhuri; 11. Timir Prakash Deb; 12. Raj Kumar Nandi; 13. Nirmal Chakravorti; 14. Ajit Kumar Chatterjee; 15. Pranab Kumar Mitra; 16. Kamal Kumar Basu.

## (d) International Statistical Education Centre (ISEC)

*Ninth term: August 1955—April 1956:*

*India:* 1. Dharnidhar Prasad; 2. Harbans Lal Chandok; 3. Md. Nooruddin; 4. Paul Jacob; 5. S. C. Jaitly;

*Pakistan:* 6. S. M. Ayub Siddique; 7. Md. Nasuruddin; 8. Mumtaz Ahmed;

*Philippines:* 9. Silverio L. Felarea; 10. Miss Rodolfo R. Madamba; 11. Francisco V. Nazarat.

## (e) Officer's Training Course : September 1955—February 1956

1. B. P. Bhargava, *Madhya Pradesh*, (9 months, National Income and Official Statistics); 2. U. C. Borah, *Assam*, (9 months, Socio-economic Surveys); 3. Abhoy Shankar Boral, *Orissa* (9 months, Machine tabulation); 4. L. N. Chaturvedi, *U.P.* (9 months, Meteorological Statistics); 5. S. P. Singh Chauhan, *Madhya Bharat*, (6 months); 6. M. N. Kaul, Anthropological Survey (6 months); 7. D. S. Kulkarni, *Madhya Pradesh*, (9 months, Socio-economic Surveys and Mechanical Tabulation); 8. M. K. Gopala-Krishnan Nair, *T.C.* (6 months); 9. Krishna Lal Narang, Ministry of Commerce, (9 months, Machine Tabulation); 10. S Narayanaswami, *Mysore*, (9 months, National Income and Regional Income Estimation); 11. D. P. Octania, *U.P.* (9 months, MT); 12. V. J. Puntambekar, *Bombay*, (9 months, Socio-economic Surveys and Mechanical Tabulation); 13. Y. Ayyappa Raju, *Andhra*, (6 months); 14. R. L. Saluja, *Delhi State*, (3 months); 15. Jagir Singh Sandhu, *Pepsu*, (9 months, Socio-economic Surveys); 16. R. Shedhani, *Madhya Bharat*, (6 months); 17. U. D. Vora, *Cutch*, (6 months); 18. R. Rangarajan, *New Delhi—Central*, (9 months, Socio-economic Surveys and Large Scale Sample Surveys).

**Appendix 11 : List of Research Scholars**

1. C. S. Ramakrishnan (*Biometric Methods*); 2. R. P. Pakshirajan (*Advanced Probability*); 3. (Mrs.) S. Nundy (*Econometrics*); 4. S. D. Chatterjee (*Stochastic Processes*); 5. B. Das Gupta (*Psychometry*); 6. Sushil Kumar (*Psychometry*); 7. T. K. Sen (*Psychometry*); 8. S. P. Sangal (*Psychometry*); 9. G. Patil (*Quality Control*); 10. D. K. Roychowdhury (*Sample Surveys*); 11. B. Bhattacharyya (*Multivariate Analysis*); 12. T. S. Varadajan (*Statistical Inference*).

**Appendix 12 : List of successful candidates in professional examinations**

Statistician's Diploma Examination : August, 1955

*A. General Papers*

*Paper I (Theoretical) :* Sasikant Dattarray Mathure (BI), Triloke Khosla (B9), Venkatesh Ranganath Kanade (B11), Dipak Chakravarti (C6), Ved Prakash Agarwal (D1), Niranjana Singh (D4), Atmaram K. Ahuja (D13), Madhusudan Shankaran Pandalai (D22), Pratap Singh Nagpaul (D25), Ramesh Shanker (L1), Jagdish Narayan Srivastava (L6), K. R. Rajagopalan (M9), D. S. Ramaratnam (M10), Arwind Parasuram Joag (P1), Narayan Kalyan Ugar (P3), Vasant Trimbak Jumde (P14), Sarasgopal Moghe (P18), Anant Raghavendra Kulkarni (P19).



## TWENTYFOURTH ANNUAL REPORT : 1955-56

*Paper II (Theoretical) :* Gopaldas S. Monga (B4), K. Mukundan (B13), Brij Nandan Pandey (C1), Chandu Lal Gupta (D5), Avtar Singh Chawla (D17), Pratap Singh Nagpaul (D25), Ramesh Shanker (L1), Jagdish Narayan Srivastava (L6), Vasant Trimbak Jumde (P14).

*Paper III (Theoretical) :* Vishnupad Srinivas Gururaj (B3), A. Gandevarian (C5), Sudhindra Nath Ganguli (D9), Narayan Kalyan Ugar (P3).

*Paper VI (Practical) :* Vedprakash Agarwal (D1), N. P. Mahadevan (M6).

*Paper VII (Practical) :* Gopaldas S. Monga (B4), Narayan Kondaji Sonavane (P2), Vasant Trimbak Jumde (P14), Anant Raghavendra Kulkarni (P19).

### B. Special Papers

#### *Papers IV & V (Theoretical)*

(1) *Statistical Quality Control :* Vasudeo Vyasacharya Galsasi (B10), Narayan Krishnaji Chandekar (C3), Dipak Chakravarti (C6), S. Rangachary (M5), Arwind Parasuram Joag (P1), Narayan Kalyan Ugar (P3), Narayan Narasinha Koti (P15), Sisirkumar Sridhar Jagdeo (P17).

(2) *Actuarial Statistics :* S. Viswanath (B15).

(3) *Mathematical Theory of Sampling Distribution :* A Gandevarian (C5), J. Shiva Rao (D11).

(4) *Economic Statistics :* Narayan Krishnaji Chandekar (C3), Y. Shiva Rao (D11).

(5) *Sample Survey (Applied) :* Kalyan Kumar Das Gupta (D7), Hiralal Jain (D23), Sadashiv Dattatray Diwanji (P11).

(6) *Design of Experiment (Applied) :* Chandu Lal Gupta (D5), Charan Singh Grewal (D6), Avtar Singh Chawla (D17), Jagdish Narayan Srivastava (L6).

(7) *Sample Survey (Theoretical) :* Sadashiv Ramchandra Gokhale (P12).

(8) *Vital Statistics and Population Studies :* Ananta Raghavendra Kulkarni (P19).

#### *Papers VIII and IX (Practical)*

(1) *Design of Experiment (Construction of Design) :* Promod Kumar Gupta (B2).

(2) *Sample Survey (Applied) :* Triloke Khosla (B9), Shivendra Bahadur (L3).

(3) *Economic Statistics :* Vasudeo Vyasacharya Ghalsasi (B10), Chandu Lal Gupta (D5).

(4) *Mathematical Theory of Sampling Distribution :* A Ghandeharian (C5).

(5) *Design of Experiment (Applied) :* Chandu Lal Gupta (D5).

(6) *Probit Analysis :* Rajagopala Rangarajan (D20).

(7) *Statistical Quality Control :* Ramesh Shanker (L1), Sadashiv Ramchandra Gokhale (P12), Narayan Narasinha Koti (P15).

(8) *Sample Survey (Theoretical) :* T. Chellaswamy (P6), Sadashiv Ramchandra Gokhale (P12).

(9) *Vital Statistics and Population Studies :* T. Chellaswamy (P6), Anant Raghavendra Kulkarni (P19).

(10) *Anthropometry :* Shivendra Bahadur (L3).



## Statistician's Diploma Examination : March 1956

*A. General Papers*

*Paper I (Theoretical)* : Tilak Raj Talwar (C20), R. L. Khanna (D19), Amar Chand Sharma (D26), Dhara Singh Sharma (D35), Shivendra Bahadur (L5), Narayan Kondaji Sonavane (P2), Balachandra Mahadeo Sathye (B6).

*Part II (Theoretical)* : Dipak Chakravarty (C7), Nirmal Kanti Das Gupta (C9), Amalendu Sen Gupta (C15), Tilak Raj Talwar (C20), Y. Shiva Rao (D1), Kambhampaty Suryanarayana Sastry (D31), K. R. Rajagopalan (M11).

*Paper III (Theoretical)* : Dipak Chakravarty (C7), Ved Prakash Aggarwal (D5), Pratap Singh Nagpal (D48), Shivendra Bahadur (L5), N. P. Mahadevan (M4), Vasant Trimbak Jumde (P4).

*Paper VI (Practical)* : Avatar Singh Chawla (D3), Rajeshwar Dayal Saxena (D15), Jagdish Narain Shrivastava (L5), Narayan Kondaji Sonavane (P2), Vasant Trimbak Jumde (P4), Triloke Khosla (B8), T. Chellaswami (B11).

*Paper VII (Practical)* : Avtar Singh Chawla (D3), Charan Singh Grewal (D10), R. L. Khanna (D19), Niranjana Singh (D23), Ramesh Shankar (L1).

*B. Special Papers**Papers IV & V (Theoretical)*

- (1) *Economic Statistics* : Dipak Chakravarti (C7), Balchandra Mahadeo Sathye (B6).
- (2) *Sample Survey (Theory)* : Tarun Kumar Gupta (D2), Charan Singh Grewal (D10).
- (3) *Design of Experiments (Applied)* : Tarun Kumar Gupta (D2).
- (4) *Statistical Quality Control* : Avtar Singh Chawla (D3), Ved Prakash Aggarwal (D5).
- (5) *Sample Survey (Applied)* : Pandurang Duyaneshwar Arola (D6), Somnath Goswami (D11), Tilakraj Mahajan (D43), Ramlal Ahuja (D49).
- (6) *Actuarial Statistics* : K. Mukundan (D36).
- (7) *Stochastic Process* : Pratap Singh Nagpaul (D48).

*Papers VIII & IX (Practical)*

- (1) *Statistical Quality Control* : Dipak Chakravarti (C7), Narayan Krishnaji Chandekar (C24), Avtar Singh Chawla (D3), Vedprakash Aggarwal (D5).
- (2) *Economic Statistics* : Dipak Chakravarti (C7).
- (3) *Design of Experiments (Applied)* : Ramesh Shankar (L1).
- (4) *Sample Survey (Theory)* : Triloke Khosla (B8), K. Ramachandran (C26).

## Computer's Certificate Examination : August 1955

*Part IA, Sec. I* : Sudhir Ch. Bhowmik (C1), Nirmal Kumar Chatterjee (C2), Amal Kumar Mukherjee (C4), Akhileshwar Banerjee (C6), Animesh Singhaw (C7), Manaj Kanti Ghosh (C9), Rabindra Nath Pal (C11), Kamal Kumar Pradhan (C12), Madan Mohan Pain

## TWENTYFOURTH ANNUAL REPORT : 1955-56

(C14), Jamuna Behari Ghosh (C17), Santi Ranjan De (C19), Nikhil Ranjan Chakravarti (C20), Miss Attivilli Sita Devi (C21), Sushil Ranjan Das (C22), Shankar Bhaduri (C23), Gopeshwar Saha (C25), Mrs. Jharna Bhattacharyya (C29), Probhat Kumar Mukhopadhyaya (C30), Mahendra Nath Banerjee (C32), Prasanta Kumar Chatterjee (C36), Manoharan Dey (C39), Nityananda Chattopadhyay (C40), Sunil Kumar Chatterjee-I (C41), Paritosh Kumar Mitra (C43), Sanjit Kumar Banerjee (C46), Swadesh Ranjan RaiChowdhury (C37), Tusar Kanti Chatterjee (C51), Manik Ratan Acharya (C57), Biswanath Chakravarti (C65), Debabrata Panji (C66), Amalendu Bhusan Sengupta (C67), Pranabananda Bhaduri (C68), Manoj Kumar Guha Thakurtha (C74), Asru Guha Thakurta (C76), Manoj Chatterjee (C78), Samir Kumar Mallick (C79), Srish Chandra Basu Rai (C80), Ranjit Kumar Naha (C82), Rabindra Nath Das (C84), Dibakar Ghosh (C86), Prasun Sen (C94), Nemai Chand Dhara (C96), Arun Prosad Singha (C98), Sunirmal Bose (C99), Bimalendu Mahalanobis (C100), Premtosh Dev (C101), Prasanta Kumar Sinha (C103), Bimal Chandra Sengupta (C107), Mrs. Arati Sarkar (C110), Bimal Jyoti Sanyal (C112), Hrishikesh Roy (C118), Subodh Kumar Paul (C123), Sreenath Paul (C125), Sourendra Nath Paul (C126), Lokenath Mukherjee (C129), Rabindra Nath Mookherjee (C130), Pravash Mukherjee (C131), Badal Kumar BasuMallik (C133), Motilal Majumdar (C134), Amiya Bhusan Majumdar (C135), Prakash Chandra Kundu (C137), Samir Ranjan Guha Roy (C141), Tarak Das Ghole (C146), Biswapati Mookherjee (C132).

*Part IA, Sec. II :* Nirmal Kumar Chatterjee (C2), Sankar Bhaduri (C23), Rabindra Nath Ghosh (C27), Mahendra Nath Banerjee (C32), Sunil Kumar Chatterjee-I (C41), Sanjit Kumar Banerjee (C46), Manik Ratan Acharya (C57), Anu Ranjan Mukherjee (C63), Biswanath Chakravarti (C65), Debabrata Panji (C66), Sudhir Chandra Chakravarti (C69), Manoj Kumar Guha Thakurta (C74), Nishit Ranjan Chowdhury (C85), Kartick Chandra Das (C89), Rajendra Lal Bhuiya (C93), Bimalendu Mahalanobis (C100), Prasanta Kumar Sinha (C103), Bimal Jyoti Sanyal (C112), Kanakeshwar Roy (C116), Swaraj Kanta Paul (C122), Lokenath Mookherjee (C129), Rabindra Nath Mukherjee (C130), Pravash Mukherjee (C131), Biswapati Mukherjee (C132), Badal Kumar Basu Maullik (C133), Amiya Bhusan Majumdar (C135), Kalyan Kumar Gupta (C140), Binayendra Goswami (C143), Tarak Das Dhole (C146).

*Part IB, Sec. I:* Nirmal Kanti Das Gupta (C8), Miss Attivilli Sita Devi (C21), Santosh Kumar Bhattacharjee (C54), Amarendra Nath Dutta (C56), Prafulla Kumar Basak (C60), Bishnu Pada Paul (C61), Anuranjan Mookerjee (C63), Bishwa Nath Ghosh (C70), Subodh Kumar Mukherjee (C71), Nihar Ranjan Mukherjee (C83), Nisit Ranjan Chowdhury (C85), Radha Shyam Nath (C97), Ajoy Kumar Sarkar (C111), Sudhir Kumar Samaddar (C113), Swaraj Kanta Paul (C122), Harekrishna Paul (C124), Rabindra Nath Mookerjee (C130), Arun Kumar Maitra (C136), Kalyan Kumar Gupta (C140), Arun Kanti Ghosal (C142), Binayendra Goswami (C143).

*Part IB, Sec. II:* Chitta Ranjan Banerjee (C10), Rabindra Narayan Paul (C11), Sankar Bhaduri (C23), Sunil Kumar Chatterjee-I (C41), Amarendra Nath Dutt (C56), Anu Ranjan Mookerjee (C63), Nishit Ranjan Chowdhury (C85), Sudhir Kumar Samaddar (C113).

*Part IC, Sec. I:* Kalayan Kumar Gupta (C140).

*Part IC, Sec. II:* Nirmal Kumar Chatterjee (C2), Nirmal Kanti Das Gupta (C8), Chittaranjan Banerjee (C10), Prasanta Kumar Chatterjee (C36), Sukhendu Moitra (C52), Miss Gita Shaha (C53), Amarendra Nath Dutt (C56), Bishnu Pada Paul (C62), Bishwanath



Ghosh (C70), Ranajit Kumar Naha (C82), Rabindra Nath Das (C84), Nishit Ranjan Chowdhury (C85), Dhiraj Mohan Sen Gupta (C108), Subodh Kumar Paul (C123), Kalyan Kumar Gupta (C140).

Computer's Certificate Examination : February 1956

*Part 1A, Sec. I:* Dhiraj Lal Rai Choudhury (C4), Ajit Kumar Chatterjee (C8), Pronab Kumar Mitra (C9), Asoke Kumar Das Gupta (C12), Mihir Kumar Rakshit (C19), Ajit Kumar Sen Gupta (C21), Panchanon Mascharak (C27), Shivendra Nath Sen Gupta (C48), Sanat Kumar Bhose (C50), Manindra Narayan Choudhury (C52), Rabindra Kumar Chakravorti (C57), Aparesh Chakravorty (C61), Amalendu Choudhury (C68), Keshab Chandra Banerjee (C74), Sanat Kumar Das (C88), Sibpada Chakravorty (C95), Lakshmi Narayan Das (C96), Sadananda Banerjee (C99), Debabrata Chandra (C108), Ramchandra Neogi (C110), Samarendra Barua (C112), Manoranjan Das (C113), Samir Baran Das Gupta (C121), Guru Narayan Samanta (C129), Kamal Kumar Basu (C130), Rajendra Lal Bhuiya (C131), Kshirode Behari Ray (C133), Timir Prakash Deb (C142), Biswadan Chattopadhyay (C156), Dhruba Ranjan Chakravorti (C157), Manindra Nath Khaskel (C169), Parimal Kumar Roy (C180), Kalidas Neogy (C181), Sukha Sinchan Roy (C191), Sunil Narayan Bose (C196), Bhagat Singh (D1), Sakti Kumar Chowdhury (G17), Debabrata Bardhan (G23).

*Part 1A, Sec. II:* Dhiraj Lal Rai Choudhury (C4), Ajit Kumar Chatterjee (C8), Pronab Kumar Mitra (C9), Asoke Kumar Das Gupta (C12), Santi Priya Bhowmik (C14), Tapan Das Gupta (C18), Mihir Kumar Rakshit (C19), Banaj Kanti Ghosh (C31), Ranjan Kumar Bhattacharyya (C35), Srish Chandra Basu Rai (C37), Sanat Kumar Bhose (C50), Nihar Ranjan Mukherjee (C51), Rabindra Kumar Chakravorti (C57), Suchendra Sekhar Das (C65), Ramendra Nath Roy (C66), Rabindra Nath Das (C72), Gopeswar Saha (C75), Subodh Kumar Mukherjee (C83), Tinkari Pal (C84), Ranajit Kumar Naha (C89), Sibpada Chakravorti (C95), Sambhunath Bhattacharjee (C97), Prasanta Kumar Chatterjee (C103), Debabrata Chandra (C108), Samir Ranjan Guha Roy (C115), Mati Lal Majumdar (C116), Samir Baran Das Gupta (C121), Guru Narayan Samanta (C129), Biswadan Chattopadhyay (C156), Dhruba Ranjan Chakravorti (C157), Amiya Kishore Das Gupta (C166), Sukha Sinchan Roy (C191).

*Part 1B, Sec. I:* Pabitra Kumar Das (C2), Mihir Kumar Rakshit (C19), Dhiraj Mohan Sen Gupta (C20), Ajit Kumar Sen Gupta (C21), Ranjan Kumar Bhattacharyya (C35), Srish Chandra Basu Rai (C37), Sanat Kumar Bhose (C50), Rabindra Kumar Chakravorti (C57), Sanjit Kumar Banerjee (C58), Biswapati Mukherjee (C86), Monoj Kumar Guha Thakurta (C132), Gour Chandra Mukherjee (C153), Bimal Jyoti Sanyal (C171), Bimalendu Mahalanobis (C194), Mohd. Yusuf Ansari (G30).

*Part 1B, Sec. II:* Kalyan Kumar Gupta (C5), Dhiraj Mohan Sengupta (C20), Ranjan Kumar Bhattacharyya (C35), Nirmal Kanti Das Gupta (C42), Sanjit Kumar Banerjee (C58), Sudhendra Sekhar Das (C65), Santosh Kumar Bhattacharjee (C78), Sibpada Chakravorty (C95), Timir Prakash Deb (C142), Gour Chandra Mukherjee (C163), Dhruba Ranjan Chakravorti (C157).

*Part 1C, Sec. I:* Sukumar Roy Choudhury (C3), Asoke Kumar Gupta (C11), Chittaranjan Dey (C17), Dhiraj Mohan Sengupta (C20), Chittaranjan Banerjee (C22), Nirmal Chandra Dey (C25), Ranjan Kumar Bhattacharyya (C35), Nirmal Kanti Das Gupta (C42), Geeta Saha (C43), Sabita Chakravorti (C49), Birendra Kumar Nandi (C67), Santosh Kumar Bhatta-



charyya (C78), Amarendra Nath Dutt (C80), Nishit Ranjan Choudhury (C82), Subodh Kumar Mukherjee (38C), Pabitra Kumar Dey Sarkar (C91), Profulla Kumar Basak (C109), Mukti Nath Mukherjee (C115), Radhashyam Nath (C160), Subimal Kanti Majumder (C190).

*Part 1C, Sec. II:* Santosh Kumar Rai Choudhuri (C1), Pabitra Kumar Das (C2), Asoke Kumar Gupta (C11), Chittaranjan Dey (C17), Nirmal Chandra Dey (C25), Tarani Kanta Paul Roy (C29), Ranjan Kumar Bhattacharyya (C35), Kalipada Chakravorti (C36), Sankar Bhaduri (C41), Soobimal Chandra Ghosh (C45), Nihar Ranjan Mukherjee (C51), Anuranjan Mukherjee (C56), Arun Kumar Maitra (C59), Sudhendra Sekhar Das (C65), Birendra Kumar Nandi (C67), Harekrishna Paul (C377), Subodh Kumar Mukherjee (C83), Purnendu Bhusan Home Roy (C98), Profulla Kumar Basak (C109), Nirmalendu Basu Choudhury (C134), Bishnu Charan Poddar (C159), Radhashyam Nath (C160), Bimal Jyoti Sanyal (C171), Nagendra Chandra Das (C172), Rabindra Nath Mukherjee (C177).

#### Statistical Field Survey Certificate Examination—August 1955

*Part 1, Sec. A:* Abhijit Rou Choudhury (C14), Lalit Mohan Chatterjee (C22), Kamalendra Nath Dutta Roy (C31), Debesh Prasanna Sen (C43), Lekraj Munzil (D1), Ramesh Chandra (D3), Sadhulal Srivastava (D6), Surinder Kumar Gandhi (D19), Anil Chandra Bhattacharyya (D21), Lakhan Lal Chaurishya (D35), Manohar Singh Babel (D36), Kailash Shankar Tandon (D40), Premnath Kapoor (G1), Debiprosad Sinha (G2), Jay Krishna Prosad (G5), Narayan Singh (G6), Tej Paul Sharma (N6), N. Wamanarao Despande (N7), M. R. Kulkarni (N8), P. N. Sahasrabudhe (N21), D. K. Ganibhir (N22), Gunendra Krishna Barori (P1), G. P. Sinha (P23), Balbir Singh Bakshi (P24).

*Part 1, Sec. B:* Anantlal Banerjee (C9), Lalit Mohan Chatterjee (C22), Kamalendra Nath Dutta Roy (C31), Lekraj Munzil (D1), Sadhulal Srivastava (D6), Surinder Kumar Gandhi (D19), Mukul Beharilal Gandhi (D25), Monohar Singh Babel (D36), Kailash Shankar Tandon (D40), Jay Krishna Prosad (G5), Narayan Singh (G6), Harinarayan Mangalamurti (N1), R. M. Bhumralkar (N14), H. Z. Fulzeli (N20), P. N. Sahasrabudhe (N21), M. S. Wate (N26), Gunendra Krishna Barori (P1), G. P. Sinha (P23).

*Part 1, Sec. C:* Kamalendra Nath Dutta Roy (C31), Ramesh Chandra (D3), Sadhulal Srivastava (D6), Surinder Kumar Gandhi (D19), Mukul Beharilal Gandhi (D25), Lakhan Lal Chaurishya (D35), Manohar Singh Babel (D36), Kailash Shankar Tandon (D40), Gunendra Krishna Barori (P1), Balbir Singh Bakshi (P24).

*Part II, Sec. A:* M. S. Rangaswamy (B8), S. V. Srinivasachar (B9), Francis Mathew Alloor (B18), S. Oliver Francis (B19), P. D. George (B20), N. S. Murty (B24), Lekraj Munzil (D1), Haken Singh Panjete (D2), D. R. Chowla (D14), Surinder Kumar Gandhi (D19), Lakhan Lal Chaurishya (D35), Manohar Singh Babel (D36), Kailash Shankar Tandon (D40), J. S. Rakhi (D39), Paranjape Sadashiv Hanumanta Rao (N27), Narayan Yeshwentrao Gore (N28), Satish Chandra Ghosh (P23), Sadananda Jha (P13), Nalini Ranjan Jha (P16), Mahesh Chandra Prosad Singh (P18), Ramnarayan Gupta (P21), Nagendra Prosad Singh (P22), Balbir Singh Bakshi (P24).

*Part II, Sec. B:* Paranjape Sadashiv Hanumanta Rao (N27), Nalini Ranjan Jha (P16), Ramnarayan Gupta (P21), C. Prabhakar (B1), S. A. Rajendran (B3), S. V. Srinivasachar (B9), Francis Mathew Alloor (B18), P. D. George (B20), N. S. Murty (B24),

Ramesh Chandra (D3), Banki Behari Mehra (D5), Bharat Bhusan Behl (D22), Manohar Singh Babel (D26), Paranjape Sadashiv Hanumanta Rao (N27), Nalini Ranjan Jha (P16), Ramnarayan Gupta (P21).

### Field Survey Certificate Examination—February 1956

*Part I, Sec. A:* Adhir Chandra Adhikari (C9), Jagadindu Sarkar (C10), Nalini Mohan Chakravorti (C12), Santosh Kumar Chatterjee (C21), Ananta Lal Banerjee (C28), Moti Singh Rathore (D2), Shadi Lall (D5), C. A. K. Arthur (D22), T. R. Sharma (D23), Amarendra Nath Mandal (G1), Ranjit Kumar Mukherjee (G6), H. S. Sachdeva (L4), D. C. Chaturvedi (L5), Kundan Lal (L8), Ram Adhar Singh (L10), Panna Lal Pal (L11), G. T. Samnani (L13), R. S. Dikshit (L18), Surendra Singh Chauhan (P3), Achintya Kumar Chatterjee (P6).

*Part I, Sec. B:* J. S. Patnaik (B2), V. Srinivasa Rao (B3), Khagendra Mohan Ganguly (C2), Nalini Mohan Chakravorti (C12), Abhijit Roy Choudhuri (C19), Rampada Dutta (C30), Sharad Madhav Wathare (D1), Moti Singh Rathore (D2), Shadi Lall (D5), Lakhan Lal Chaurishya (D19), C. A. K. Arthur (D22), T. R. Sharma (D23), Ramesh Chandra (D24), Surinder Nath Kapur (D27), Amarendra Nath Mondal (G1), Premnath Kapoor (G2), Thakur Indradeo Sharma (G5), Ranjit Kumar Mukherjee (G6), H. S. Sachdeva (L4), D. C. Chaturvedi (L5), Kundan Lal (L8), Panna Lal Pal (L11), G. T. Samnani (L13), R. S. Dikshit (L18), Premprakash Sharma (P2), Surendra Singh Chauhan (P3), Kameshwar Nath Srivastava (P5), P. K. Kalkarni (PN4).

*Part I, Sec. C:* V. Srinivasa Rao (B3), Khagendra Mohan Ganguly (C2), Santi Ranjan Nandi (C4), Sudhir Chandra Chakravorti (C6), Motilal Bhattacharyya (C40), Sharad Madhav Wathare (D1), Moti Singh Rathore (D2), Lekhraj Munzil (D4), Shadi Lall (D5), C. A. K. Arthur (D22), T. R. Sharma (D23), Prem Nath Kapoor (G2), Jai Krishna Prasad (G3), Narayan Singh (G9), H. S. Sachdeva (L4), D. C. Chaturvedi (L5), Kundan Lal (L8), Ram Adhar Singh (L10), Panna Lal Pal (L11), G. T. Samnani (L13), R. S. Dikshit (L18), Prem Prakash Sarma (P2), Surendra Singh Chauhan (P3), Kameshwar Nath Srivastava (P5), Achintya Kumar Chatterjee (P6), P. K. Kulkarni (PN4), Nilkanth Wamanrao Deshpande (X23).

*Part II, Sec. A:* R. Seshadri (B1), S. R. David (B7), G. K. Venugopal (B9), S. A. Rajendran (B10), P. Ramkrishna Menon (B11), C. Prabhakar (B16), Moti Singh Rathore (D2), Shadi Lall (D5), Bharat Bhusan Behl (D7), Kailash Chandra Chopra (D12), Yogendra Paul Seth (D13), Bankey Behari Mehra (D17), Ram Lal Ahuja (D18), Brehma Prakash Soni (D20), C. A. K. Arthur (D22), Ramesh Chandra (D24), Hukum Chand Chaurishya (D25), Bhola Nath Roy (L1), S. N. Srivastava (L2), M. Ranjan Ali (L3), H. S. Sachdeva (L4), S. S. Banga (L9), S. S. Yadava (L21), Surendra Singh Chauhan (P3), Y. D. Shende (PN2), Z. B. Kothari (PN21).

*Part II, Sec. B:* R. Seshadri (B1), S. R. David (B7), S. Oliver Francis (B8), P. Ramkrishna Menon (B11), Agni Kumar Das (C29), Joharlala Ghosh (C32), Lekh Raj Munzil (D4), Shadi Lall (D5), Kailash Chandra Chopra (D12), Lakhan Lal Chaurishya (D19), Bhola Nath Roy (L1), M. Ramzan Ali (L3), H. S. Sachdeva (L4), S. S. Banga (L9), Godadhar Prasad (P10), Z. B. Kothari (PN21).



# SANKHYĀ

## THE INDIAN JOURNAL OF STATISTICS

*Edited by : P. C. MAHALANOBIS*

VOL. 17, PART 4

FEBRUARY

1957

### SOME OBSERVATIONS ON INPUT-OUTPUT ANALYSIS

*By* OSKAR LANGE

*School of Planning and Statistics, Warsaw*

#### 1. THE SCOPE OF INPUT-OUTPUT ANALYSIS

The analysis of inter-industry relations, usually referred to as input-output analysis, serves the purpose of establishing the quantitative relations between various branches of production which must be maintained in order to assure a smooth flow of production in the national economy. It studies the conditions of mutual consistency of the outputs of the various branches of the national economy which result from the fact that the output of one branch is the source of input in other branches.

The idea that certain proportions must be maintained between the outputs of various branches of the national economy is at the basis of the equilibrium analysis of classical political economy and neo-classical economics. The proportions referred to are, however, conceived by classical and neo-classical economic theory basically in 'horizontal' terms, i.e., as proportions between final products designed to satisfy the wants of consumers. Under conditions of competitive capitalism, of free mobility of capital, the tendency of the rate of profit towards a 'normal' level in each branch of the national economy leads towards an equilibrium of output of the various branches. In equilibrium, output is adjusted to the demand for the various products. In a planned economy, it is believed, proper planning should assure the establishment of equilibrium proportions.

While this idea of 'horizontal' equilibrium proportions undoubtedly points to an important aspect of the relations between the output of the various branches of the national economy, it overlooks the need of maintaining another kind of proportions, determined not by conditions of consumers' demand, but by conditions of technological relations associated with the fact that the output of certain products serves



—entirely or in part—as input in the process of producing other products. We may call this the problem of ‘vertical’ proportions.

This problem of ‘vertical’ proportions is the subject matter of input-output analysis. The problem was first posed by Quesnay in his famous ‘Tableau Economique’. Its insight was lost by classical and neo-classical economic theory. A systematic treatment as well as the fundamental solution of the problem was given by Marx in his schemes of reproduction of capital contained in volume II of *Das Kapital*. Outside of Marxist political economy the problem was scarcely seen, neo-classical economics confining itself to the study of equilibrium conditions of the ‘horizontal’ type.

However, in business cycle theory of bourgeois economists the problem of ‘vertical relations’ between investment goods and consumers’ goods was bound to reappear, for it is this type of relation which is at the bottom of the phenomenon of crises and depressions. Consequently, it plays an important role in Keynesian theory. The ‘vertical’ character of the relations involved causes that ‘disproportionalities’ in this field are not automatically solved by the process of competition through capital moving from less profitable to more profitable branches of the economy. It also explains why smooth economic development is not automatically assured under conditions of capitalism, even independently of the handicaps resulting from the specific features of monopoly capitalism.

The importance of a study of the ‘vertical’ relations between various branches of the economy, *i.e.* of input-output analysis, is not limited to conditions of a capitalist economy. As was already pointed out by Marx, since input-output relations are based on technological conditions of production, proper proportions in this field must be maintained in any economic system. A study of such relations is therefore necessary for purposes of socialist economic planning as well as for the understanding of the working-mechanism of capitalist economy. Under conditions of socialism input-output analysis is a necessary tool of ascertaining the internal consistency of national economic plans.

In the socialist countries input-output analysis takes the form of various ‘statistical balances’ which serve as tools of national economic planning. These balances are conceived as concretisations of the general idea underlying the reproduction schemes of Marx. In the USA Professor Leontief has developed a type of input-output analysis which, too, can be conceived as a concretisation of Marx’s idea of input-output relations taking place in the process of reproduction of the national product. Professor Leontief’s analysis takes explicitly into account the technological relations between output and input. Though applied first to the economy of the USA, this analysis like all input-output analyses is also applicable to a socialist economy. Indeed, it seems to me, that this analysis achieves its full justification only if applied as a tool of economic planning. Its technique, though first applied to a capitalist economy, points beyond the historical limitations of capitalism and can come fully into its own only under conditions of planned economy.

# SOME OBSERVATIONS ON INPUT-OUTPUT ANALYSIS

## 2. THE MARXIAN SCHEMES

Marx's analysis of reproduction is based on two devices. First, the value of the total national product during a period of time (e.g. a year) is considered as being composed of three parts—the value of the means of production used-up during this period (to be denoted by  $c$ —in Marx's terminology the constant capital used up), the value of the labour power directly engaged in production (to be denoted by  $v$ —in Marx's terminology the variable capital, i.e., the revolving wage fund), the surplus generated (to be denoted by  $s$ ). Thus:

Total national product =  $c+v+s$ .

Here,  $c$  is the replacement of the means of production used-up,  $v+s$  is the total value added (or national income).

Secondly, the national economy is divided into two departments: one producing means of production, the other producing consumers' goods. Using the subscripts 1 and 2 to indicate the two departments, respectively, we shall write:

$$\begin{aligned}\text{total output of means of production} &= c_1 + v_1 + s_1 \\ \text{total output of consumers' goods} &= c_2 + v_2 + s_2 \\ \text{total national product} &= c + v + s\end{aligned}$$

where

$$c = c_1 + c_2, \quad v = v_1 + v_2, \quad s = s_1 + s_2.$$

In a stationary economy (Marx's simple reproduction) :

$$\begin{aligned}\text{total demand for means of production} &= c_1 + c_2 \\ \text{total demand for consumers' goods} &= v_1 + v_2 + s_1 + s_2\end{aligned}$$

The total demand for means of production is equal to the joint replacement requirement of both departments, the total demand for consumers' goods is equal to the joint wage fund and surplus of both departments.

Putting equal demand and output of means of production, we obtain

$$c_1 + c_2 = c_1 + v_1 + s_1 \quad \dots (2.1)$$

which simplifies to

$$c_2 = v_1 + s_1. \quad \dots (2.2)$$

The same result is obtained from putting equal total demand and output of consumers' goods.

That is

$$v_1 + v_2 + s_1 + s_2 = c_2 + v_2 + s_2. \quad \dots (2.3)$$

This is so, because the total national product  $c+v+s$  is being given. Equation (2.3) can be deduced from equation (2.1).

Equation (2.2) indicates an input-output relation between the two departments of the national economy. Indeed, let us write,

$$\begin{aligned}& \left[ \begin{array}{c} c_1 \\ c_2 \end{array} \right] + \left[ \begin{array}{c} v_1 + s_1 \\ v_2 + s_2 \end{array} \right] \\ & \dots (2.4)\end{aligned}$$

Department 1 produces means of production. Part of its output equal in value to  $c_1$  is retained within the department for replacement of the means of production used up. The remainder (in the rectangle) equal in value to  $v_1 + s_1$  is transmitted to department 2 in exchange for consumers' goods. Department 2 produces consumers' goods. Part of its output equal in value to  $v_2 + s_2$  is retained within the department for consumption. The remainder in the rectangle equal in value to  $c_2$  is transmitted to department 1 in exchange for the means of production needed for replacement of those which were used-up. In order that production goes on smoothly, the output of the two departments must be co-ordinated in such a way that a balanced exchange takes place between the two departments, i.e.,  $c_2 = v_1 + s_1$ . The above table (2.4) thus indicates the input output relations between the two departments: equation (2.2) gives the condition of proper balance between the two departments.

In an expanding economy (Marx's expanded reproduction) not all the surplus is consumed; part of it is accumulated to increase the amount of means of production and to employ more labour power. We shall express this by writing,

$$s = \bar{s} + s_c + s_v$$

where  $\bar{s}$  is the part of the surplus consumed,  $s_c$  the part of the surplus used to increase the amount of means of production,  $s_v$  the part of the surplus used to employ more labour power.

Dividing as before, the economy into two departments, we have,

$$\text{total output of means of production} = c_1 + v_1 + \bar{s}_1 + s_{1c} + s_{1v}$$

$$\text{total output of consumers' goods} = c_2 + v_2 + \bar{s}_2 + s_{2c} + s_{2v}$$

$$\text{total national product} = c + v + \bar{s} + s_c + s_v$$

Furthermore;

$$\text{total demand for means of production} = c_1 + c_2 + s_{1c} + s_{2c}$$

$$\text{total demand for consumers' goods} = v_1 + v_2 + s_{1v} + s_{2v} + \bar{s}_1 + \bar{s}_2$$

The total demand for means of production is equal to the joint replacement and expansion requirement of both departments. The total demand of consumers' goods is equal to the joint wage fund, the joint expansion of the wage fund and the joint surplus consumed in both departments.

Equality of demand and output of means of production implies

$$c_1 + s_{1c} + c_2 + s_{2c} = c_1 + v_1 + \bar{s}_1 + s_{1c} + s_{1v} \quad \dots (2.5)$$

which leads to

$$c_2 + s_{2c} = v_1 + \bar{s}_1 + s_{1v} \quad \dots (2.6)$$

The same result can be obtained from the condition of equality of demand and output of consumer's goods.



## SOME OBSERVATIONS ON INPUT-OUTPUT ANALYSIS

Equation (2.6) indicates the input-output relation between the two departments in an expanding economy. It can be presented by means of the following table :

$$\begin{array}{c|c} c_1 + s_{1c} & + v_1 + \bar{s}_1 + s_{1v} \\ \hline c_2 + s_{2c} & + v_2 + \bar{s}_2 + s_{2v} \end{array} \quad \dots \quad (2.7)$$

In department 1 part of the product equal in value to  $c_1 + s_{1c}$  is retained within the department for replacement of the means of production used up and for expansion of the amount of means of production in the department. The remainder (contained in the rectangle) is transmitted to department 2 in exchange for consumers' goods. In department 2 part of the product equal in value to  $v_2 + \bar{s}_2 + s_{2v}$  is retained for consumption. The remainder (contained in the rectangle) is transmitted to department 1 in exchange for means of production for replacement of the means of production used-up and for expansion of the amount of means of production in the department. The proper balance between the two departments is thus expressed by equation (2.6).

### 3. INPUT-OUTPUT RELATIONS IN A MULTI-SECTOR MODEL

Professor Leontief's input-output tables are designed to study the relations between a larger number of sectors of the national economy. Let the economy be divided into  $n$  production sectors denoted by the indices 1, 2, ...,  $n$ . Denote by  $X_i$  the total or gross output of the  $i$ -th sector by  $X_{ij}$  the quantity of the product of the  $i$ -th sector transmitted to the  $j$ -th sector where it is used as input. Further denote by  $x_i$  the net output of the  $i$ -th sector, viz., that part of the gross output  $X_i$  which is not allocated to another sectors to be used there as input. The net output  $x_i$  can be consumed, exported, or accumulated for the purpose of investment.

We have thus,

$$X_i = \sum_{j=1}^n X_{ij} + x_i \quad (i = 1, 2, \dots, n). \quad \dots \quad (3.1)$$

It is convenient to represent the input-output relations between the sectors of the economy in the form of a table as follows:

$X_1$	$X_{11}$	$X_{12}$	...	...	...	...	$X_{1n}$	$x_1$
$X_2$	$X_{21}$	$X_{22}$	...	...	...	...	$X_{2n}$	$x_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$X_n$	$X_{n1}$	$X_{n2}$	...	...	...	...	$X_{nn}$	$x_n$

... (3.2)

The items in the square matrix in the center of the table represent the input-output relations, or the 'interflows' between the various branches of the national

economy (also called 'intersector deliveries'.) The column on the right hand side represents the net outputs and the column on the left hand side the gross outputs of the various products. The rows are subject to the balance relation indicated by equation (3.1).

Since the process of production requires not only the use of means of production but also the application of direct labour, we may supplement the above input-output table by introducing the amounts of labour force employed in production. Let us denote the total labour force available in the national economy by  $X_0$  the labour force employed in producing the output of the  $i$ -th sector of the economy by  $X_{0i}$  and, finally, by  $x_0$  the labour force not employed productively. The latter may be either unemployed (labour reserve) or employed in non-productive occupations, i.e., in occupations which do not produce material goods (e.g., personal services). With regard to the allocation of the total labour force the following equation holds :

$$X_0 = \sum_{i=1}^n X_{i0} + x_0. \quad \dots (3.3)$$

Introducing the allocation of the labour force into the input-output table, we obtain the following table

$X_0$	$X_{01}$	$X_{02}$	...	...	...	...	$X_{0n}$	$x_0$
$X_1$	$X_{11}$	$X_{12}$	...	...	...	...	$X_{1n}$	$x_1$
$X_2$	$X_{21}$	$X_{22}$	...	...	...	...	$X_{2n}$	$x_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$X_n$	$X_{n1}$	$X_{n2}$	...	...	...	...	$X_{nn}$	$x_n$
	$Y_1$	$Y_2$	...	...	...	...	$Y_n$	

... (3.4)

The items in the square matrix in the center of the table are 'interflows' for 'inter-sector deliveries'. The upper row in the center represents the allocation of the labour force to the various branches of the economy. Similarly as before, the column at the right represents the remainder of the labour force not allocated productively ( $x_0$ ), and the net outputs of the various products ( $x_i$ ;  $i = 1, \dots, n$ ). The column on the left hand side represents the total labour force  $X_0$  and the gross outputs  $X_i$  ( $i = 1, 2, \dots, n$ ) of the various branches.

The entries in table may be expressed either in physical units or in value units. In the latter case, the table is sometimes called a 'transaction table' rather than our input-output table. Whatever the units, the rows of the table can always be summed, for each row is expressed in the same units (e.g., man-hours, tons, gallons, yards, pieces). Thus the equations (3.1) and (3.2) hold under all circumstances. We may call them the 'allocation equations'.

# SOME OBSERVATIONS ON INPUT-OUTPUT ANALYSIS

The columns, however, can be summed only if the entries of the table are expressed in value units (e.g., rupees) i.e. if the table is a transaction table. otherwise the items of a column would be non-homogeneous. We shall write these sums in the following form.

$$Y_j = X_{0j} + \sum_{i=1}^n X_{ij} \quad (j = 1, 2, \dots, n). \quad \dots \quad (3.5)$$

Obviously,  $Y_j$  is the cost of the output of the  $j$ -th branch,  $X_{0j}$  being the cost of the labour force employed and  $\sum X_{ij}$  the cost of the means of production used-up in producing the output. We may call the equations (3.5) the 'cost equations'. The costs of producing the output of the various branches of the economy are indicated in the row at the bottom of table (3.4).

The excess of the value of the output of a branch of the national economy over the cost of producing the output is the surplus produced in this branch. Denoting the surplus produced in the  $j$ -th branch by  $S_j$ , we have

$$S_j = X_j - Y_j \quad \dots \quad (3.6)$$

and in view of (3.5), 
$$X_j = X_{0j} + \sum_{i=1}^n X_{ij} + S_j \quad (j = 1, \dots, n). \quad \dots \quad (3.7)$$

This is the relation which in a multi-sector model corresponds to the Marxian decomposition of the value of the output of a branch of the national economy into  $c_j + v_j + s_j$  ( $j = 1, 2$ ). Here  $\sum X_{ij}$  stands for  $c_j$  and  $X_{0j}$  stands for  $v_j$  in the Marxian notation. The value added in the sector is  $X_{0j} + S_j$ .

Introducing the surplus produced in the various branches of the economy into the transaction table and taking account of the relation (3.7) we obtain the following transaction table:

$X_0$	$X_{01}$	$X_{02}$	.....	$X_{0n}$	$x_0$
$X_1$	$X_{11}$	$X_{12}$	.....	$X_{1n}$	$x_1$
$\ddots$	$\ddots$	$\ddots$	.....	$\ddots$	$\ddots$
$X_n$	$X_{n1}$	$X_{n2}$	.....	$X_{nn}$	$x_n$
	$S_1$	$S_2$	.....	$S_n$	
	$X_1$	$X_2$	.....	$X_n$	

... (3.8)

From table (3.8) it is apparent that the gross output of a branch, say  $X_i$ , can be obtained either by summation of the entries of a row or by summation of the entries of a column. Consequently, we have

$$\sum_{j=1}^n X_{ij} + x_i = X_{0i} + \sum_{j=1}^n X_{ji} + S_i \quad (i = 1, \dots, n). \quad \dots \quad (3.9)$$



This results directly from the equations (3.1) and (3.7). On both sides of equation (3.9)  $X_{ii}$  is appearing under the summation sign : it is the part of the output retained in the sector for replacement. Eliminating  $X_{ii}$  from the equation, we obtain

$$\sum_{j \neq i} X_{ij} + x_i = X_{0i} + \sum_{j \neq i} X_{ji} + S_i \quad (i = 1, \dots, n). \quad \dots (3.10)$$

This equation states that (measured in value units) the outflow from the sector to other sectors—plus the net output is equal to the inflow from other sectors plus the value added in the sector.

Equation (3.10) is the analogue, in a multisector model, of the Marxian equations (3.2) and (3.6) of the previous section which hold in a two-sector model. The mentioned Marxian equations are obtained—just like equation (3.10)—by putting equal the value of the output of the sector and the total allocation of the sector's output and by eliminating on both sides the part of the output retained in the sector.

In order to see the exact analogy of equation (3.10) and the equations of the Marxian two-sector model, let us transform equation (3.10) in the following way. Suppose that the net output  $x_i$  is partly reinvested in the sector and partly consumed or allocated to other sectors; the corresponding parts will be indicated by  $x'_i$  and  $x''$  respectively. Thus we have

$$x_i = x'_i + x''_i \quad (i = 1, \dots, n). \quad \dots (3.11)$$

Further, suppose that the surplus produced in the sector is used partly for consumption, partly for employment of additional labour force in the sector, and partly for addition to the means of production used in the sector. Denote these quantities by  $S_i$ ,  $S_{i0}$  and  $x'_i$  respectively. Thus

$$S_i = \bar{S}_i + S_{i0} + x'_i. \quad \dots (3.12)$$

Substituting (3.11) and (3.12) into equation (3.10) and eliminating  $x'_i$  on both sides, the equation reduces to

$$\sum_{j \neq i} X_{ij} + x''_i = \sum_{j \neq i} X_{ji} + X_{0i} + S_{i0} + \bar{S}_i \quad (i = 1, \dots, n). \quad \dots (3.13)$$

In this form not only the quantities  $X_{ii}$  retained in the sector for replacement but also the quantity retained in the sector for expansion is eliminated. Equation (3.13) states that the net outflow to other sectors and to consumption is equal to the inflow from other sectors and to the part of the value added not retained in the sector. This is the exact counterpart—in a multisector model—to the Marxian equation (3.6) in the previous section.

If the number of sectors is reduced to two, equation (3.13) becomes identical with equation (3.2) of the preceding section. In this case (3.13) reduces to

$$X_{12} + x''_1 = X_{21} + X_{01} + S_{10} + \bar{S}_1. \quad \dots (3.14)$$

# SOME OBSERVATIONS ON INPUT-OUTPUT ANALYSIS

The corresponding transaction table takes the form:

$X_0$	$X_{01}$	$X_{02}$	$x'_{01} + x'_{02} + x''_0$
$X_1$	$X_{11}$	$X_{12}$	$x'_1 + x''_1$
$X_2$	$X_{21}$	$X_{22}$	$x'_2 + x''_2$
	$\bar{S}_1$ $S_{10}$ $x'_1$	$\bar{S}_2$ $S_{20}$ $x'_2$	

... (3.15)

Sector 1 produces means of production, sector 2 produces consumers' goods. As consumer's goods are not a means of production,  $X_{21} = 0$ , and as means of production are not consumed,  $x''_1$  are the means of production allocated to sector 2 for expansion. Using the notation of the preceding section, we shall write:

$$\begin{aligned} X_{01} &= v_1; & X_{02} &= v_2 \\ X_{11} &= c_1; & X_{12} &= c_2; & X_{21} &= 0 \\ x''_2 &= s_{2c}; & S_{10} &= s_{1v} \end{aligned}$$

Thus equation (3.15) takes the form

$$c_2 + s_{2c} = v_1 + s_{1v} + \bar{s}_1$$

which is identical with equation (2.6) of the preceding section. In a stationary economy,  $s_{2c} = s_{1v} = 0$ , and the equation reduces to  $c_2 = v_1 + s_1$ , i.e., to equation (2.2) of the preceding section.

It should also be noticed that of the equations (3.10) or (3.13) (which are equivalent to (3.10)), only  $n-1$  are independent. From the transaction table (3.8) it is apparent that

$$\sum_i (\sum_j X_{ij} + x_i) \equiv \sum_i (X_{0i} + \sum_j X_{ji} + S_i) \equiv \sum_i X_i \quad \dots (3.16)$$

This implies directly that one of the equations (3.10) can be deduced from the remaining  $n-1$ . This corresponds to the property of the Marxian two sector model where only one relation like equation (2.6) or (2.2) of the preceding section holds between the two sectors.

Eliminating the double sums on both sides of the identity (3.16), we obtain

$$\sum_i x_i = \sum_i X_{0i} + \sum_i S_i \quad \dots (3.17)$$

which indicates that the net product of the national economy, or national income is equal to the total value added during the period under consideration.

## 4. TECHNOLOGICAL RELATIONS AND VALUE RELATIONS

In order to study the effect of the technological conditions of production upon input-output relations we have to distinguish sharply between input-output tables expressed in physical units and transaction tables which are expressed in value units. For this purpose we shall use a separate notation.

The physical output of the  $i$ -th sector will be denoted by  $Q_i$ , the physical net output by  $q_i$  and the physical interflow from the  $i$ -th to the  $j$ -th sector by  $q_{ij}$  ( $i, j = 1, \dots, n$ ). The physical total labour force (measured, for instance, in properly weighted man-hours) will be denoted by  $Q_0$ , the physical labour power employed in the  $i$ -th sector by  $q_{0i}$  and the remainder not employed productively by  $q_0$ . The physical input-output table can thus be written in the form

$Q_0$	$q_{01}$	$q_{02}$	.....	$q_{0n}$	$q_0$
$Q_1$	$q_{11}$	$q_{12}$	.....	$q_{1n}$	$q_1$
$Q_2$	$q_{21}$	$q_{22}$	.....	$q_{2n}$	$q_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$Q_n$	$q_{n1}$	$q_{n2}$	.....	$q_{nn}$	$q_n$

... (4.1)

The rows of the table are subject to the allocation balance

$$Q_i = \sum_j q_{ij} + q_i \quad (i = 0, 1, 2, \dots, n). \quad \dots (4.2)$$

The technological conditions of production can be described by the technical coefficients, called also coefficients of production:

$$a_{ij} = q_{ij}/Q_j \quad (i = 0, 1, \dots, n; j = 1, \dots, n). \quad \dots (4.3)$$

The coefficient  $a_{0j}$  indicates the labour power employed in producing a unit of output of the  $j$ -th sector, the remaining coefficients  $a_{ij}$  indicate the amount of output of the  $i$ -th sector needed to produce a unit of output of the  $j$ -th sector.

In the socialist countries the values of these coefficients are generally available in form of the 'technical norms' used in planning and administration of production. These norms indicate the amounts of labour power, raw materials etc., which are allowed to be used per unit of output. In the absence of such 'technical norms' in the industries the technical coefficients can be obtained approximately from statistical input-output tables, according to formula (4.3). This method was employed by Professor Leontief.

Introducing the technical coefficients (4.3), the allocation equations (4.2) become

$$Q_i = \sum_j a_{ij} Q_j + q_i \quad (i = 0, 1, \dots, n).$$

It is convenient to separate the first equation relating to labour power from the remaining ones. We have then

$$Q_0 = \sum_j a_{0j} Q_j + q_0 \quad \dots (4.4)$$

and the remaining equation can be written in the form

$$(1 - a_{ii})Q_i - \sum_{j \neq i} a_{ij} Q_j = q_i \quad (i = 1, \dots, n). \quad \dots (4.5)$$



# SOME OBSERVATIONS ON INPUT-OUTPUT ANALYSIS

Thus the equations (4.5) can be solved separately from equation (4.1). The matrix of the coefficients of these equations

$$\begin{pmatrix} 1-a_{11}, -a_{12} & \dots & \dots & -a_{1n} \\ \dots & \dots & \dots & \dots \\ -a_{n1}, -a_{n2}, & \dots & \dots & 1-a_{nn} \end{pmatrix} \dots \quad (4.6)$$

is called the 'technical matrix'. It describes the technological conditions of production.<sup>1</sup>

In the system (4.5) there are  $n$  equations and  $2n$  variables, i.e., the gross outputs  $Q_1, \dots, Q_n$  and the net outputs,  $q_1, \dots, q_n$ . If the technical matrix is non-singular as we shall assume to be the case, there are thus  $n$  degrees of freedom. We can fix in the national economic plan the net outputs  $q_1, \dots, q_n$  and the gross outputs  $Q_1, \dots, Q_n$  are then uniquely determined by the equations (4.5). Or, instead, we can fix in the plan the gross outputs and the net outputs available which will result uniquely from the equations. Or, finally, we can fix in the plan a number of gross outputs and of net outputs, together  $n$  in number — and the remaining  $n$  gross and net outputs are determined by the equations.

If the technical matrix happens to be singular, the number of degrees of freedom is increased according to the order of nullity of the matrix. Thus if the rank of the matrix is  $m$  ( $m < n$ ), the order of nullity is  $n-m$  and the number of degrees of freedom is  $n + n-m$ . Thus we must fix in the plan  $2n-m$  variables, the remaining  $m$  variables being then obtained from the equations (4.5).

Having the gross outputs  $Q_1, \dots, Q_n$  either from the equations (4.5) or directly from the plan, we can substitute them into equation (4.4). This gives us the total labour force employed  $\sum_{j=1}^n a_{0j}Q_j$ , and taking the total labour force  $Q_0$  as a datum, we can calculate  $q_0$  i.e., the labour force remaining outside productive employment.

To show the relation between the transaction table and the physical input-output table (1), we must take explicitly account of prices. Denote by  $p_0$  the remuneration of a unit of labour force, and by  $p_1, p_2, \dots, p_n$  the prices of the products of the various sectors. Further  $p'_0$  denotes the earning of the labour force not employed in production. We have then

$$\begin{aligned} X_i &= p_i Q_i, \quad x_i = p_i q_i & \dots & (4.7) \\ x_0 &= p'_0 q_0, \\ X_{ij} &= p_i q_{ij}. \end{aligned}$$

<sup>1</sup> It should be noticed that this technical matrix differs from the matrix used by Professor Leontief in so far that in Professor Leontief's matrix the coefficients  $a_{ii}$  in the diagonal are absent; his diagonal consists only of unities. This is due to the fact that he does not take into account the fact that part of the output is retained in the sector as means of production, e.g., part of the output of agriculture is retained as seed and as fodder for breeding of animals, part of the coal is retained in the coal mines on fuel etc. If the number of sectors in the model is small, the sectors being accordingly large, this omission may be serious.

We shall also denote by  $\Pi_i$  the surplus per unit of gross physical output of the sector, i.e.,

$$S_i = \Pi_i Q_i \quad (i = 1, \dots, n). \quad \dots (4.8)$$

Introducing these relations into the transaction table (4.8) of the preceding section we obtain the following form of the transaction table:

$p_0 \sum q_{0j} + p'_0 q_0$	$p_0 q_{01},$	$p_0 q_{02},$	.....,	$p_0 q_{0n}$	$p'_0 q_0$
$p_1 Q_1$	$p_1 q_{11}$	$p_1 q_{12},$	.....,	$p_1 q_{1n}$	$p_1 q_1$
$p_2 Q_2$	$p_2 q_{21}$	$p_2 q_{22},$	.....,	$p_2 q_{2n}$	$p_2 q_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots \vdots \vdots$	$\vdots$	$\vdots$
$p_n Q_n$	$p_n q_{n1}$	$p_n q_{n2},$	.....,	$p_n q_{nn}$	$p_n q_n$
	$\Pi_1 Q_1$	$\Pi_2 Q_2$	.....,	$\Pi_n Q_n$	
	$p_1 Q_1$	$p_2 Q_2$	.....,	$p_n Q_n$	

... (4.9)

Summing the columns we obtain the equations

$$p_0 q_{0i} + \sum_j p_j q_{ji} + \Pi_i Q_i = p_i Q_i$$

which are identical with equations (3.7) in the preceding section. Taking account of the technical coefficients ( $a_{ij}$ ), these equations can be written.

$$a_{0i} p_0 + \sum_j a_{ji} p_j + \Pi_i = p_i$$

or, more conveniently,  $(1 - a_{ii}) p_i - \sum_{j \neq i} a_{ji} p_j - a_{0i} p_0 = \Pi_i \quad \dots (4.10)$

The matrix of the coefficients is

$$\begin{pmatrix} 1 - a_{11}, & -a_{21}, & \dots & \dots & -a_{n1}, & -a_{01} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ -a_{1n}, & -a_{2n}, & \dots & \dots, & 1 - a_{nn}, & -a_{0n} \end{pmatrix} \quad \dots (4.11)$$

There are  $n$  equations and  $2n+1$  variables i.e.,  $n$  prices  $p_1, \dots, p_n$  the wage rate  $p_0$  and  $n$  per-unit surpluses,  $\Pi_1, \dots, \Pi_n$ . If the matrix is of rank  $n$ , there are thus  $n+1$  degrees of freedom. We can fix, for instance, the wage rate  $p_0$  and the per unit surpluses  $\Pi_1, \dots, \Pi_n$ , the  $n$  prices are then uniquely determined. Or, instead, we can fix the  $n$  prices mentioned and the wage rate, the per unit surpluses are then uniquely determined, or any other combination of  $n+1$  variables can be fixed, the  $n$  remaining ones resulting from the equations.

# SOME OBSERVATIONS ON INPUT-OUTPUT ANALYSIS

If the rank of the matrix is less than  $n$ , the number of degrees of freedom increases correspondingly. The important point to be noticed is that these relations between prices of products, wage rate and per unit surpluses are entirely determined by the technological conditions of production as represented by the technical matrix of the coefficients of equations (4.10). The  $n \times n$  submatrix containing the first  $n$  columns is simply the transpose of the technical matrix (4.6).

Now we can show the relation between the physical input-output relations and the input-output relations in value terms as expressed in a transaction table. The rows of the transaction table (4.9) are subject to the allocation balance

$$p_i Q_i = \sum_j p_i q_{ij} + p_i q_i$$

or, introducing the technical coefficients according to (4.3)

$$p_i Q_i = \sum_j p_i a_{ij} Q_j + p_i q_i$$

This can also be written in the form

$$p_i Q_i = \sum_j a'_{ij} p_j Q_j + p_i q_i \quad \dots (4.12)$$

$$\text{where} \quad a'_{ij} = (p_i/p_j) a_{ij} \quad (i, j = 1, \dots, n). \quad \dots (4.13)$$

In view of (4.7), the equations (4.12) can be written in the form

$$X_i = \sum_j a'_{ij} X_j + x_i$$

$$\text{or} \quad (1 - a'_{ii}) X_i + \sum_{j \neq i} a'_{ij} X_j = x_i \quad (i = 1, \dots, n). \quad \dots (4.14)$$

These equations establish the relations between the value of the net outputs  $x_1, \dots, x_n$  and the value of the gross outputs of the various sectors.

The matrix of the coefficients of these equations is

$$\begin{pmatrix} 1 - a'_{11} & -a'_{12} & \dots & \dots & \dots & -a'_{1n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ -a'_{n1} & -a'_{n2} & \dots & \dots & \dots & 1 - a'_{nn} \end{pmatrix} \quad \dots (4.15)$$

i.e., analogous to the matrix (4.6), only that the coefficients  $a'_{ij}$  appear instead of the coefficients  $a_{ij}$ .

The coefficients  $a'_{ij}$  can be written in the form

$$a'_{ij} = X_{ij}/X_j \quad (i, j = 1, \dots, n). \quad \dots (4.16)$$

They indicate the value of the input of the product of the  $i$ -th sector ( $i = 1, \dots, n$ ) required to produce a unit of value of output of the  $j$ -th sector. We shall call these coefficients the 'input coefficients'.



In addition, input coefficients of the type

$$a_{oj} = X'_{oj}/X_j \quad \dots (4.17)$$

can be introduced which indicate the value of direct labour power needed to produce a unit of value of product of the  $j$ -th sector. With the aid of these coefficients the value of the total labour force employed in production can be calculated, i.e.,

$$X_0 - x_0 = \sum_j a'_{oj} X_j. \quad \dots (4.18)$$

The input coefficients derive their significance from their simple behaviour with regard to aggregation of two or several sectors into one single sector. For instance, let us aggregate the  $j$ -th sector and the  $k$ -th sector and denote the new sector thus obtained as the  $l$ -th sector.

The value of the gross output of the new sector is then

$$X_l = X_j + X_k \quad \dots (4.19)$$

and the value of the part of the product of the  $i$ -th sector allocated as input to the new sector is

$$X_{il} = X_{ij} + X_{ik} \quad \dots (4.20)$$

The new input coefficient is, consequently,

$$a'_{il} = X_{il}/X_l = \frac{X_{ij} + X_{ik}}{X_j + X_k}.$$

In view of the definition (4.16), this is equal to

$$a'_{il} = \frac{a'_{ij} X_j + a'_{ik} X_k}{X_j + X_k} \quad \dots (4.21)$$

i.e., the new input coefficient is the weighted mean of the input coefficients before aggregation.

The input coefficients can be given a simple interpretation on the basis of the Marxian theory of value. If the prices of the products express the amount of socially necessary labour required to produce a physical unit of output, the input coefficients indicate the quantity of social labour engaged in one sector necessary to produce in another sector a unit of value (i.e., an amount representing a unit of social labour.) This quantity is entirely determined by the technological conditions of production. The transaction table indicates the allocation of the social labour among the various sectors of the national economy and shows the interflow of social labour between the various sectors of the economy. Aggregation of sectors can be performed by mere summation and the input coefficients are transformed under aggregation by simple averaging.

## SOME OBSERVATIONS ON INPUT-OUTPUT ANALYSIS

The Marxian theory, however, points out that in a capitalist economy prices do not exactly reflect the amount of social labour necessary to produce a unit of output. Systematic deviations arise between the 'prices of production', i.e., equilibrium prices under competitive capitalism, and the values of products measured in labour. These deviations are the result of the technologically determined differences in ratios of capital goods and direct labour employed on one hand, and the equalisation of the rates of profit by competition on the other hand. Monopoly produces further systematic deviations. Consequently, transaction tables of a capitalist economy give only an approximate picture of allocation of social labour. In a socialist economy transaction tables give a picture of the allocation of social labour to the extent that prices express the amount of social labour required in production. Therefore, in a socialist economy, a proper system of prices reflecting the amounts of social labour required in production is a necessary instrument of effective accounting of the allocation of society's labour force among the various branches of national economy.

### 5. CONSUMPTION AND INVESTMENT

The net output of any sector of the national economy may be consumed, exported or accumulated for future use. Accumulated output may be designed for future consumption or allocated to increase the quantity of means of production, i.e., invested in the process of production. In the first case we shall consider it as another form of consumption; the last mentioned use will be called productive investment. The part of the net output exported can be considered as destined for consumption or productive investment in proportion as the goods imported in return consist of consumers' goods or means of production. Thus the total net output of a sector may be divided up into a part consumed and a part utilized for productive investment.

Consider the net physical output  $q_i$  of the  $i$ -th sector and denote the part consumed by  $q_i^{(1)}$  and the part invested productively by  $q_i^{(2)}$ . Then

$$q_i = q_i^{(1)} + q_i^{(2)}. \quad \dots (5.1)$$

Further

$$k_i = q_i^{(1)}/Q_i; \quad \alpha_i = q_i^{(2)}/Q_i \quad \dots (5.2)$$

Thus,  $k_i$  is the proportion of the gross output  $Q_i$  of the sector  $i$  consumed, and  $\alpha_i$  is the proportion of the gross output  $Q_i$  used for productive investment. We shall call them the 'rate of consumption' and 'rate of investment', respectively.

Obviously,

$$q_i = (k_i + \alpha_i) Q_i \quad \dots (5.3)$$

The allocation equations (4.5) of the preceding section can then be written as homogeneous equations of the form

$$(1 - a_{ii} - k_i - \alpha_i)Q_i - \sum_{j \neq i} a_{ij} Q_j = 0 \quad (i = 1, \dots, n). \quad \dots (5.4)$$

In order that these have a non-trivial solution it is necessary that

$$\begin{vmatrix} 1-a_{11}-k_1-\alpha_1 & -a_{12} & \dots & \dots & -a_{1n} \\ \dots & \dots & \dots & \dots & \dots \\ -a_{n1} & -a_{n2} & \dots & \dots & 1-a_{nn}-k_n-\alpha_n \end{vmatrix} = 0 \quad (5.5)$$

i.e., the rates of consumption and rates of investment of the various sectors cannot be fixed independently of each other. Their mutual relations depend on the rank of the matrix of (5.5).

This may be conveniently illustrated by the example of a two sector model. Taking the sectors 1 and 2, the determinantal equation (5.4) becomes

$$(1-a_{11}-k_1-\alpha_1)(1-a_{22}-k_2-\alpha_2) = a_{12}a_{21} \quad (5.6)$$

or, 
$$\frac{1-a_{11}-k_1-\alpha_1}{a_{12}} = \frac{a_{21}}{1-a_{22}-k_2-\alpha_2} \quad (5.7)$$

This means that the fractions of the gross output of each sector going to the other sector for current use in production, i.e.,  $1-a_{ii}-k_i-\alpha_i$  is proportional to the technical co-efficients relating the two sectors to each other. It is seen from (5.6) that if the rates of consumption are kept constant, the rate of investment of one sector can be increased only at the expense of reducing the rate of investment of the other sector. A similar relation holds for the rates of consumption of the two sectors, if the rates of investment are kept constant.

Now suppose that sector 1 produces means of production and sector 2 produces consumers' goods. Means of production are needed to produce consumers' goods but themselves are not consumed; consequently,  $a_{12} > 0$  and  $k_1 = 0$ . Consumers' goods are only usable for consumption; they are neither needed currently to produce means of production nor are they investable in production. Consequently,  $a_{21} = 0$  and  $\alpha_2 = 0$ . Thus the equation (5.6) turns into

$$(1-a_{11}-\alpha_1)(1-a_{22}-k_2) = 0$$

As consumers' goods are not invested, their total net output is consumed, i.e.,  $1-a_{22}-k_2 = 0$ . Consequently,  $1-a_{11}-\alpha_1$  is arbitrary and the rate of investment  $\alpha_1$ , can be arbitrarily fixed.

In a communist economy distribution of the national product is divorced from the input of labour and follows the principle, 'to each according to his need'. Under such circumstances, the rates of consumption can be set by policy provided their mutual relations resulting from (5.5) are observed. These relations are entirely expressed in physical terms and no value relations are involved; they depend entirely on the technical coefficients.



## SOME OBSERVATIONS ON INPUT-OUTPUT ANALYSIS

In a socialist economy distribution of the national product is based on the remuneration for labour performed. Under capitalism it depends also on property in means of production which permits certain classes to appropriate the surplus generated in production. Therefore, in a socialist economy the rates of consumption are related to the remuneration of the labour force both in productive and non-productive employment. In a capitalist economy they depend also on the use property owners make of the surplus they appropriate.

In order to determine the rates of consumption, it is best to start from a transaction table. We have seen in section 3, equation (3.17), that the net product of the national economy is equal to the total value added in production, i.e.,

$$\sum_i x_i = \sum_i X_{0i} + \sum_i S_i.$$

Introducing the rates of consumption and of investment, we can write this in the form

$$\sum_i k_i X_i = \sum_i X_{0i} + \sum_i S_i - \sum_i \alpha_i X_i. \quad \dots (5.8)$$

The left hand side of this equation represents the part of the total value of the net product of the economy (national income) devoted to consumption.

Let  $W_i$  be the fraction of the part of the national income devoted to consumption spent for the product of the  $i$ -th sector ( $i = 1, \dots, n$ ). We consider these fractions to be 'behavioural data' and shall call them 'consumption parameters'. Then

$$k_i X_i = W_i (\sum_j X_{0j} + \sum_j S_j - \sum_j \alpha_j X_j), \quad (i = 1, \dots, n; \quad \sum W_i = 1). \quad \dots (5.9)$$

(The subscripts in the summation signs on the right hand side are denoted by  $j$  in order to avoid confusion with the subscript  $i$  on the left hand side).

Introducing input coefficients and writing

$$S_j = \Pi'_j X_j \quad (j = 1, \dots, n) \quad \dots (5.10)$$

we can write

$$k_i X_i = W_i (\sum_j a'_{0j} X_j + \sum_j \Pi'_j X_j - \sum_j \alpha_j X_j) \quad (i = 1, \dots, n). \quad \dots (5.11)$$

Substituting this in the allocation equations (4.14) of the preceding section which indicate the allocation balances in the rows of the transaction table, we obtain

$$[1 - a'_{ii} - \alpha_i - W_i(a'_{0i} + \Pi'_i - \alpha_i)]X_i - \sum_{j \neq i} [a'_{ij} + W_i(a'_{ij} + \Pi'_i - \alpha_j)]X_j = 0, \quad \dots (5.12)$$

$$(i = 1, \dots, n).$$

In order that these equations have a non-trivial solution we must have the determinant

$$\begin{vmatrix} 1-a'_{11}-\alpha_1-W_1(a'_{01}+\Pi'_1-\alpha_1) & \dots & -a'_{1n}-W_1(a'_{0n}+\Pi'_n-\alpha_n) \\ \dots & \dots & \dots \\ -a'_{n1}-W_n(a'_{01}+\Pi'_1-\alpha_1) & \dots & 1-a'_{nn}-\alpha_n-W_n(a'_{0n}+\Pi'_n-\alpha_n) \end{vmatrix} = 0. \quad (5.13)$$

This condition establishes the relations which must be maintained between the rates of investment  $\alpha_1, \dots, \alpha_n$  when the rates of consumption are determined by the 'demand equations' (5.1).

The expressions

$$a'_{0j} + \Pi'_j - \alpha_j \quad (j = 1, \dots, n) \quad \dots (5.14)$$

which occur in the determinant (5.5) indicate the part of the value added per unit of output value of the sector which is devoted to consumption. By multiplying these expressions by  $W_i$  we get the fraction of it which goes into consumption of the product of the  $i$ -th sector.

For illustration let us consider a two sector model. The determinantal equation can then be written in the form

$$\frac{1-a'_{11}-\alpha_1-W_1(a'_{01}+\Pi'_1-\alpha_1)}{a'_{12}+W_1(a'_{02}+\Pi'_2-\alpha_2)} = \frac{a'_{21}+W_2(a'_{01}+\Pi'_1-\alpha_1)}{1-a'_{22}-\alpha_2-W_2(a'_{02}+\Pi'_2-\alpha_2)} \quad \dots (5.15)$$

This equation indicates that the fraction of the value of gross output of each sector remaining after deduction of the part retained in the sector for replacement ( $a'_{ii}$ ), and for consumption  $W_i(a'_{0i}+\Pi'_i-\alpha_i)$  and of the part devoted to investment ( $\alpha_i$ ) is proportional to the total demand (per unit of value of its output) of the other sector for the product of the first. The latter is equal to the sum of the input coefficient— $a'_{ij}$  and the output of the other sector required for consumption, i.e.,  $W_i(a'_{0j}+\Pi'_j-\alpha_j)$ .

Transforming the input coefficients into technical coefficients according to formula (4.13) of the preceding section and observing that

$$\Pi'_j = \frac{\Pi_j}{p_j}, \quad (j = 1, \dots, n) \quad \dots (5.16)$$

we can write the determinantal equation (5.13) in the abbreviated form

$$\left| \delta_{ij} - \frac{p_i}{p_j} a_{ij} - W_i \left( \frac{p_0}{p_j} a_{0j} + \frac{\Pi_j}{p_j} - \alpha_j \right) \right| = 0 \quad \dots (5.17)$$

## SOME OBSERVATIONS ON INPUT-OUTPUT ANALYSIS

where  $\delta_{ij} = 1$  for  $i = j$  and  $\delta_{ij} = 0$  for  $i \neq j$ . This equation contains the wage rate  $p_0$ , the product prices  $p_1, \dots, p_n$  and the per-unit surpluses  $\Pi_1, \dots, \Pi_n$ . These quantities cannot be eliminated from the equation.

Thus when the rates of consumption are determined by 'demand equations' like (5.11) linking them to the national income, the relation between the rates of investment in the various sectors of the national economy cannot be expressed in purely physical and technological terms. They have to be expressed in value terms and are found according to (5.13) to depend on the input coefficients, the rates of surplus  $\Pi'_1, \dots, \Pi'_n$  and the consumption parameters  $W_1, \dots, W_n$  of the various sectors.

As in the light of the Marxian theory of value the input coefficients can be interpreted as indicating technological conditions of production, the relations between the rates of investment are found to depend, in addition to the technological conditions of production, on behavioural parameters relating consumption of the various products to national income and on the per-unit surpluses in the various sectors. The latter can be considered as 'sociological parameters'. In a capitalist economy they are equal to the proportion of the value of each sector's output appropriated by the owners of means of production. In a socialist economy the surpluses are set by considerations of social policy, providing the resources for productive investment and for society's collective consumption.

### 6. INVESTMENT AND ECONOMIC GROWTH

The part of the net outputs of the various sectors invested in production is added to the means of production available in the next period. This makes possible in the next period an increase in the output of the various sectors of the national economy. The investment done in one period adds to the amount of means of production in operation in the next period. In consequence, a larger output is obtained in the next period. The outputs of successive period (years, for instance) are linked up in a chain through the investments undertaken in each period. Thus, productive investment generates a process of growth of output.

Let  $Q_i(t)$  be the gross physical output of the  $i$ -th sector of the economy during the time period indicated by  $t$ , e.g., the year 1955, and let  $\alpha_i$  be the rate of investment of the  $i$ -th sector as defined by (5.2) in the preceding section. The quantity of the output of the sector invested is thus  $\alpha_i Q_i(t)$ . By this amount increases the stock of product of the  $i$ -th sector available in the economy as means of production.

This increment is partly retained in the sector and partly allocated to other sectors. Denote the increment allocated to the  $j$ -th sector by  $\Delta q_{ij}(t)$ , ( $i, j = 1, \dots, n$ ). The index  $t$  indicates the period during which the allocation takes place.

$$\text{We have} \quad \alpha_i Q_i(t) = \sum_j \Delta q_{ij}(t). \quad \dots \quad (6.1)$$

However, not all the increment allocated is used-up by the various sectors during a single unit period of time. For instance, if it consists of machines or other durable equipment it will last for several units of time (years) and only a fraction of



it is used up during a unit period of time. Let the durability of the part of the output of the  $i$ -th sector allocated to the  $j$ -th sector as additional means of production be  $T_{ij}$  units of time.  $T_{ij}$  is taken as a parameter given by the technological conditions of production and may be called the 'turnover period' of the particular type of productive equipment. The reciprocal of the turnover period, i.e.,  $1/T_{ij}$  is the rate of used up per unit of time, it is also called 'rate of replacement' or 'rate of amortisation'.

In order to produce a unit of physical output of the product of the  $j$ -th sector during a unit period of time the quantity  $a_{ij}$  of the product of the  $i$ -th sector must be used up during that period of time;  $a_{ij}$  is the technical coefficient. Thus to increase in the next period the output of the  $j$ -th sector by an additional unit, the quantity of output of the  $i$ -th sector  $a_{ij} \cdot T_{ij}$  must be allocated to the  $j$ -th sector. Then exactly  $a_{ij}$  of output of the  $i$ -th sector will be used-up in the next unit period in the sector and this will produce one unit of output.

The quantities

$$b_{ij} = a_{ij}T_{ij} \quad (i, j = 1, \dots, n) \quad \dots (6.2)$$

may be called the 'investment coefficients'. The investment coefficients indicate the quantity of output of one sector which must be invested in the other sector in order to increase by one unit the other sector's output in the next unit period.

The investment coefficients as well as their reciprocals reflect technological conditions of production; given the technical coefficients, the investment coefficients are proportional to the turnover periods of the various types of means of production.

Write  $Q_j(t)$  for the physical gross output of the  $j$ -th sector in the unit period under consideration and  $Q_j(t+1)$  for the physical gross output of this sector in the next unit period. An increment of output of the  $j$ -th sector equal to  $Q_j(t+1) - Q_j(t)$  requires the investment in the sector of the following quantity of the output of  $i$ -th sector.

$$\Delta q_{ij} = b_{ij}[Q_j(t+1) - Q_j(t)] \quad (i, j = 1, \dots, n). \quad \dots (6.3)$$

In view of (6.1), we have

$$\alpha_i Q_i(t) = \sum_j b_{ij}[Q_j(t+1) - Q_j(t)] \quad (i = 1, \dots, n). \quad \dots (6.4)$$

These equations express the relations between the allocation of the part of the net product of each sector devoted to investment in the various sectors of the economy and the increments of output obtained in the various sectors in the next unit period.

If the amounts of product of the various sectors invested during the unit period  $t$ , i.e.,  $\alpha_i Q_i(t)$  are given ( $i = 1, \dots, n$ ), the increments of output in the next unit period can be calculated from the equations (6.4).

Denote by

$$B \equiv \begin{pmatrix} b_{11}, & b_{12}, & \dots, & b_{1n} \\ b_{21}, & b_{22}, & \dots, & b_{2n} \\ \dots & \dots & \dots & \dots \\ b_{n1}, & b_{n2}, & \dots, & b_{nn} \end{pmatrix} \quad \dots (6.5)$$

# SOME OBSERVATIONS ON INPUT-OUTPUT ANALYSIS

the matrix of the investment coefficients. The increments of output in the various sectors are then

$$Q_j(t+1) - Q_j(t) = \frac{1}{|B|} \sum_i |B_{ij}| \alpha_i Q_i(t) \quad \dots \quad (6.6)$$

where  $|B|$  is the determinant of the matrix  $B$  and  $|B_{ij}|$  is the co-factor of the element.

It is convenient to write

$$B_{ij} = \frac{|B_{ij}|}{|B|} \quad \dots \quad (6.7)$$

and express (6.6) in the form

$$Q_j(t+1) - Q_j(t) = \sum_i B_{ij} \alpha_i Q_i(t) \quad (j = 1, \dots, n). \quad \dots \quad (6.8)$$

The coefficients  $B_{ij}$  indicate the increment of output obtained in the  $j$ -th sector from an additional unit of the  $i$ -th sectors' product invested in the  $j$ -th sector. They may be called 'intersector output-investment ratios'. The matrix of the coefficients  $B_{ij}$  is the inverse of the matrix  $B$ .

The increments of output in the various sectors depend on the investment coefficients, and on the amounts of product of the various sectors invested. The investment coefficients, in turn, depend on the technical coefficients and turnover periods. By virtue of (6.2) the matrix of investment coefficients can be presented as follows:

$$B = \begin{pmatrix} a_{11}T_{11} & a_{12}T_{12} & \dots & a_{1n}T_{1n} \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ a_{n1}T_{n1} & a_{n2}T_{n2} & \dots & a_{nn}T_{nn} \end{pmatrix} \quad \dots \quad (6.9)$$

In this way the investments done in one unit period lead to an increase of output in the next period. If the rates of investment remain constant, the investments in the successive unit periods are,

$$\alpha_i Q_i(t+1), \alpha_i Q_i(t+2), \dots, \quad (i = 1, \dots, n).$$

The investments of the first unit period  $t$  are the initial 'shock' which sets in motion the process of economic growth. The investments in the successive unit periods carry the process forward from one stage to another.

The course of the process of economic growth can be deduced from the equation (6.4) or, for that matter, also from the equivalent equations (6.8). These are linear difference equations with constant coefficients. The characteristic equation of the system (6.4) is

$$0 = \begin{vmatrix} \alpha_1 + b_{11}(1-\lambda) & b_{12}(1-\lambda) & \dots & b_{1n}(1-\lambda) \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ b_{n1}(1-\lambda) & b_{n2}(1-\lambda) & \dots & \alpha_{nn} + b_{nn}(1-\lambda) \end{vmatrix} \quad \dots \quad (6.10)$$

The solution of the difference equations indicating the gross output in the unit period  $t_s$  can be written in the form

$$Q_j(t_s) = \sum C_k h_{jk} \lambda_k^{t_s} \quad (j = 1, \dots, n) \quad \dots (6.11)$$

where the  $\lambda_k$  are the roots of the characteristic equation, the  $C_k$  are constants determined by the outputs  $Q_j(t_s)$  in the initial unit period  $t_s$ , the  $h_{jk}$  are constants determined by the matrix of the coefficients of equation (6.4), i.e., by the matrix

$$\begin{pmatrix} \alpha_1 + b_{11}, & b_{12}, & \dots, & b_{1n} \\ \dots & \dots & \dots & \dots \\ b_{n1}, & b_{n2}, & \dots, & \alpha_n + b_{nn} \end{pmatrix} \quad \dots (6.12)$$

Thus the constants  $C_k$  reflect the initial situation of the national economy while the constants  $h_{jk}$  depend on the technological structure of the economy as expressed by the technical coefficients and the turnover periods as well as on the rates of investment.<sup>1</sup>

This analysis can be generalized by considering the rates of investment as variable in time, i.e., considering functions  $\alpha(t)$  instead of constants  $\alpha_i (i = 1, \dots, n)$ . In a similar way, changes in technical coefficients and turnover periods can be investigated. Instead of the constant investment coefficients, we would have to consider functions of time  $b_{ij}(t)$ , where  $i, j = 1, \dots, n$ . The difference equations (6.5) become then,

$$\alpha_i(t) Q_i(t) = \sum_j b_{ij}(t) [Q_j(t+1) - Q_j(t)] \quad \dots (6.13)$$

Since the coefficients in these equations are not constant, the equations require more complicated methods of treatment.

The increments in output from one unit period to the next one can, however, be easily computed. They are, in analogy with (6.8),

$$Q_j(t+1) - Q_j(t) = \sum_i B_{ij} \alpha_i(t) Q_i(t), \quad \dots (6.14)$$

the matrix of the coefficients  $B_{ij}$  being now the inverse of the matrix

$$B(t) = \begin{pmatrix} b_{11}(t), & b_{12}(t), & \dots, & b_{1n}(t) \\ \vdots & \vdots & \vdots & \vdots \\ b_{n1}(t), & b_{n2}(t), & \dots, & b_{nn}(t) \end{pmatrix} \quad \dots (6.15)$$

The relations between investment and the process of growth of output are here presented entirely in physical terms. They are found to depend solely on the techno-

<sup>1</sup> In the above, the roots  $\lambda_k$  are assumed to be all distinct. In case of a multiple root the corresponding  $h_{jk}$  on the right hand side of (6.11) is not a constant but a polynomial of degree one less than the multiplicity of the root. The coefficients of this polynomial are determined by the technological structure of the economy expressed by the matrix and the rates of investment. The coefficients  $C_k$  remain determined by the initial situation.



## SOME OBSERVATIONS ON INPUT-OUTPUT ANALYSIS

logical structure of the economy and on the rates of investment chosen. The process of economic growth, however, can also be presented in value terms.

In such a case, the technological investment coefficients  $b_{ij}$  are replaced by a set of coefficients.

$$b'_{ij} = \frac{\Delta X_{ij}}{X_j(t+1) - X_j(t)} \quad (i, j = 1, \dots, n), \quad \dots \quad (6.16)$$

indicating the value of the output of the  $i$ -th sector which must be invested in the  $j$ -th sector in order to obtain in the latter a unit increment of output value. These coefficients may be called 'investment-outlay coefficients' or simply, 'outlay coefficients'.<sup>1</sup>

In view of the relations (4.7) in section 4, the outlay coefficients are related to the investment coefficients as follows:

$$b'_{ij} = \frac{p_i}{p_j} \cdot b_{ij}. \quad \dots \quad (6.17)$$

Taking into account (6.2), they can also be written in the form:

$$b'_i = a'_{ij} T_{ij} = \frac{p_i}{p_j} \cdot a_{ij} T_{ij}. \quad \dots \quad (6.18)$$

Using the relations (4.7) of section 4 the difference equations (6.4) expressing the relations between investments in the various sectors of the economy and the increments of output obtained can be written in the value form:

$$\alpha_i X_i(t) = \sum_j b'_{ij} [X_j(t+1) - X_j(t)], \quad \dots \quad (6.19)$$

and the solutions of these equations are obtained by means of their characteristic equation which is

$$0 = \begin{vmatrix} \alpha_1 + b'_{11}(1-\lambda), & \dots & \dots & b'_{1n}(1-\lambda) \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ b'_{n1}(1-\lambda), & \dots & \dots, & \alpha_n + b'_{nn}(1-\lambda) \end{vmatrix} \quad \dots \quad (6.20)$$

The process of growth of the value of the output of the various sectors of the economy is thus determined—given the values of the initial outputs  $X_1(t_0)$ , ...,  $X_n(t_0)$  by the outlay coefficients  $b'_{ij}$  and the rates of investment  $\alpha_{ij}$ .

---

<sup>1</sup>Usually the term 'capital-coefficients' is used to denote the outlay coefficients. For reasons exposed by the Marxian theory the term 'capital' is not appropriate in a socialist economy because it covers up the fundamental difference between the role of capital as value of means of production used to appropriate by their owners the surplus produced in the national economy and the role of means of production as an instrument in the physical process of production. We, therefore, prefer to use the term 'outlay coefficients', meaning by 'outlay' the money value of the physical investments.

The outlay coefficients behave under aggregation of two or several sectors into one sector in a similar way like the input coefficients. The outlay coefficients of the new sector resulting from aggregation are the weighted means of the outlay coefficients of the sectors aggregated.

Indeed, denote by the subscript  $l$  the sector resulting from aggregation of the  $j$ -th sector and the  $k$ -th sector. The outlay coefficients of the new sector are then

$$b_u = \frac{\Delta X_u}{X_l(t+1) - X_l(t)}.$$

Since

$$\begin{cases} \Delta X_u = \Delta X_{uj} + \Delta X_{uk} \\ X_l(t) = X_j(t) + X_k(t) \\ X_l(t+1) = X_j(t+1) + X_k(t+1) \end{cases} \quad \dots \quad (6.21)$$

we obtain, taking into account the definition (6.16),

$$b'_u = \frac{b'_{uj} [X_j(t+1) - X_j(t)] + b'_{uk} [X_k(t+1) - X_k(t)]}{[X_j(t+1) - X_j(t)] + [X_k(t+1) - X_k(t)]} \quad \dots \quad (6.22)$$

The merit of presentation of the process of growth of output resulting from investment in value terms consists in the possibility it gives to aggregate sectors. But it must be pointed out that the outlay coefficients do not reflect only the technological structure of the economy. As seen from (6.17), they depend also on the relative prices of the products. The result of their averaging under aggregation also depends on the relative prices of the products of the sectors aggregated.

However, on the basis of the Marxian theory of value, the outlay coefficients may, under appropriate circumstances, be interpreted as indicating the quantity of social labour employed in the sector of the economy which must be 'stored up' in order to increase the output of another by an amount representing one unit of social labour. Under such interpretation, which requires that prices reflect the amounts of social labour necessary to produce a physical unit of product, the outlay coefficients too represent the technological structure of the economy.

The way in which the growth of output set in motion by investment depends entirely on the technological structure of the economy is further elucidated by the fact that the investment coefficients are, according to (6.2), products of the technical coefficients and the turnover periods, or that the outlay coefficients, according to (6.8) are the products of the input-coefficients and the turnover periods.<sup>1</sup> Thus the technological conditions determining the growth of output resulting from investment

---

\*The fact that the investment coefficient are not independent of the technical coefficients but are derived from them by multiplication by the turnover periods seems to have been pointed out first by David Hawkins, "Some conditions of macroeconomic stability," *Econometrica* 1948, p. 313. Usually they are wrongly taken as independent data, like for instance by Professor Leontief in, *Studies in the Structure of the American Economy*, Oxford University Press, New York 1953, p. 56.

## SOME OBSERVATIONS ON INPUT-OUTPUT ANALYSIS

consist entirely of two factors. One are the technical coefficients indicating current input-output relations during a unit period. The other are the turnover periods which simply indicate the durability of the various means of production and, consequently the rate of use-up of the means of production in a single unit period of time.

This disposes definitely of any mystical notions about the 'productivity' of a mythical entity 'capital' conceived as a separate factor of production distinguished from the physical means of production. Such metaphysical entity is proved to be non-existent.

In a capitalist economy 'capital' consists of private property rights to means of production which permit the owners of the means of production to appropriate the surplus produced in the national economy. 'Capital' is the power to appropriate surplus. This power, under capitalism, is measured by the money value of the means of production and hired labour power a person (or corporation) can command. In a socialist economy such property rights are absent. There exist simply physical means of production and certain technological conditions expressed by the technical coefficients and turnover periods. From these technological conditions there result certain consequences concerning the quantity of social labour which must be 'stored up' in order to achieve a planned increase in output. Thus there is no need in a socialist economy for any concept of 'capital'. Such concept would only obscure the technological character of the conditions of the process of economic growth.

### 7. EFFECTS OF INVESTMENT ON NATIONAL INCOME AND EMPLOYMENT

The equations (6.19) of the preceding section can be transformed in a shape analogous to equation (6.8), i.e., in a shape which presents the increment of the value of output of a sector of the national economy as a linear combination of the investments undertaken in the various sectors. For greater generality it is convenient to consider the rates of investment,  $\alpha_i$ , as variable in time, i.e.,  $\alpha_i(t)$ . We obtain then,

$$X_j(t+1) - X_j(t) = \sum_i B'_{ij} \alpha_i(t) X_i(t) \quad (j = 1, \dots, n). \quad \dots (7.1)$$

The coefficients  $B'_{ij}$  are the elements of a matrix  $(B_{ij})^{-1}$  which is the inverse of the matrix of the outlay coefficients

$$B' \equiv \begin{pmatrix} b'_{11} & b'_{12} & \dots & \dots & b'_{1n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ b'_{n1} & b'_{n2} & \dots & \dots & b'_{nn} \end{pmatrix} \quad \dots (7.2)$$

This means that,

$$B'_{ij} = \frac{|B'_{ij}|}{|B'|}; \quad (i, j = 1, \dots, n), \quad \dots (7.3)$$

where  $|B'|$  is the determinant of  $B$  and  $|B'_{ij}|$  is the co-factor of the element  $b'_{ij}$ .



The coefficients  $B'_{ij}$  may be called 'intersector output-outlay ratios'. They indicate the increment of the output (measured in value) of the  $j$ -th sector resulting from a unit increase of investment outlay in the  $i$ -th sector.

Summing the equation (7.1) over all sectors of the national economy, we obtain

$$\sum_j [X_j(t+1) - X_j(t)] = \sum_j \sum_i B'_{ij} \alpha_i(t) X_i(t)$$

or, writing

$$\beta_i = \sum_j B'_{ij} \quad (i = 1, \dots, n). \quad \dots (7.4)$$

$$\sum_j (X_j(t+1) - X_j(t)) = \sum_i \beta_i \alpha_i(t) X_i(t). \quad \dots (7.5)$$

The left hand side of equation (7.5) is the increment, from one unit period to the next, of gross national product. The coefficients  $\beta_i$  on the right hand side indicate the effect of a unit increase in investment outlay in the various sectors of the economy on national gross product. They can be called simply 'output-outlay ratios' of the various sectors.

A further simplification of equation (7.5) can be achieved by expressing the investment outlays in the various sectors as a fraction of the total investment outlay in the national economy. Denote by  $\alpha(t)$  the overall rate of investment in the national economy during the unit period  $t$ . The total investment outlay during the unit period is

$$\alpha(t) \sum_i X_i(t).$$

Denoting further by  $\mu_i(t)$  the proportion of the total investment outlay which is undertaken in the  $i$ -th sector of the economy, we have

$$\begin{aligned} \alpha_i(t) X_i(t) &= \mu_i(t) \alpha(t) \sum_i X_i(t); \\ (\sum_i \mu_i(t) &= 1). \end{aligned} \quad \dots (7.6)$$

Substituting the relation (7.6) into equation (7.5) and observing that

$$\sum_i X_i(t) = \sum_j X_j(t),$$

we arrive at

$$\sum_j (X_j(t+1) - X_j(t)) = \alpha(t) \sum_j X_j(t) \sum_i \beta_i \mu_i(t),$$

which also can be written as

$$\frac{\sum_j (X_j(t+1) - X_j(t))}{\sum_j X_j(t)} = \alpha(t) \sum_i \beta_i \mu_i(t). \quad \dots (7.7)$$

The left hand side of (7.7) is the rate of increase of gross national product and will be denoted by  $r(t)$ . In order to simplify the right hand side we shall put

$$\beta(t) = \sum_i \beta_i \mu_i(t) \quad \dots (7.8)$$

## SOME OBSERVATIONS ON INPUT-OUTPUT ANALYSIS

Since  $\sum_i \mu_i(t) = 1$ ,  $\beta$  can be interpreted as the average output-outlay ratio of the national economy. Equation (7.7) can thus be expressed in the simple form

$$r(t) = \alpha(t) \beta(t). \quad \dots (7.9)$$

Thus the rate of increase of gross national product is the product of the overall rate of investment and of the average output-outlay ratio.

Now we can calculate the effect of a given investment programme upon gross national income after a number of unit periods of time. Let  $\sum_j X_j(t_0)$  be the gross national product in the initial unit period  $t_0$ , and let the investment programme be given by the overall rates of investment  $\alpha(t_0), \dots, \alpha(t_n)$  and the fractions  $\mu_i(t_0), \dots, \mu_i(t_n)$  of the total investment outlay allocated to the various sectors of the economy, ( $i = 1, \dots, n$ ). We obtain, then, the average output-outlay ratios,  $\beta(t_0), \dots, \beta(t_n)$ . The gross national product in unit period  $t_s$  ( $t_s > t_0$ ) is,

$$\sum_j X_j(t_s) = \prod_{t=t_0}^{t_s} [1 + \alpha(t)\beta(t)] \sum_j X_j(t_0). \quad \dots (7.10)$$

If the overall rate of investment  $\alpha(t)$  and the allocation fractions  $\mu_i(t)$  are the same during each unit period, say  $\alpha$  and  $\mu_i$ , this reduces to

$$\sum_j X_j(t_s) = (1 + \alpha\beta)^{t_s - t_0} \cdot \sum_j X_j(t_0). \quad \dots (7.11)$$

National income is the value of the total net output of the national economy. The value of the net output of the  $i$ -th sector in unit period,  $t$  is according to the allocation equation (4.12) or (4.14)

$$x_i(t) = X_i(t) - \sum_j a'_{ij} X_j(t), \quad \dots (7.12)$$

where the  $a'_{ij}$  are input coefficients. National income in unit period  $t$  thus is

$$\sum_i x_i(t) = \sum_i X_i(t) - \sum_j X_j(t) \sum_i a'_{ij}.$$

Remembering that  $\sum_i X_i(t) = \sum_j X_j(t)$

$$\sum_i x_i(t) = \sum_j x_j(t)$$

we obtain

$$\sum_j x_j(t) = (1 - \sum_{i,j} a'_{ij}) \sum_j X_j(t). \quad \dots (7.13)$$

Thus the national income in any unit period differs from the gross national product of that period by a constant factor,  $(1 - \sum_{i,j} a'_{ij})$ . The double sum in this factor expresses the fraction of national product which is allocated for replacement of the products used up in the process of production during the unit period (i.e. for replace-

ment). The factor itself indicates the fraction of gross national products which constitutes net product, i.e., national income.

Since national income differs from gross national product by a constant multiplier, the rate of increase of national income is necessarily equal to the rate of increase of gross national product. Consequently, the relation (7.9) holds for national income as well as gross national product.

Furthermore, we find that national income in unit period  $t_s$  is related to national income in the initial unit period  $t_0$  ( $t_s > t_0$ ) by formulae analogous to (7.10) and (7.11), namely,

$$\sum_j x_j(t_s) = \prod_{t=t_0}^{t_s} [1 + \alpha(t) \beta(t)] \sum_j x_j(t_0), \quad \dots (7.14)$$

and, in the case when  $\alpha(t) = \text{const}$  and  $\beta(t) = \text{const}$

$$\sum_j x_j(t_s) = (1 + \alpha\beta)^{t_s - t_0} \sum_j x_j(t_0). \quad \dots (7.15)$$

The total employment generated by the gross national product is calculated as follows. Denote, as in section 4 by  $a'_{oj}$  the input coefficient indicating the value of direct labour force needed to produce a unit of value of product in the  $j$ -th sector. We shall call them for convenience 'employment coefficients'. The total employment (in value units) corresponding to gross national product in unit period  $t$  is, according to the balance equation (4.1)

$$\sum_j a'_{oj} X_j(t).$$

Consequently, the increment of total employment from one unit period to the next is  $\sum_j a'_{oj} [X_j(t+1) - X_j(t)]$ .

Taking into account equation (7.1), we find

$$\sum_j a'_{oj} [X_j(t+1) - X_j(t)] = \sum_j a'_{oj} \sum_i B'_{ij} \alpha_i(t) X_i(t),$$

or, in view (7.6),

$$\sum_j a'_{oj} [X_j(t+1) - X_j(t)] = \sum_j a'_{oj} \sum_i B'_{ij} \mu_i(t) \alpha(t) \sum_i X_i(t). \quad \dots (7.16)$$

This expression can be simplified as follows. Write

$$\gamma_i = \sum_j a'_{oj} B'_{ij} \quad (i = 1, \dots, n), \quad \dots (7.17)$$

$\gamma_i$  is the additional amount of employment (in value units) created in the national economy by a unit increase in investment outlay in the  $i$ -th sector of the economy. We may call it the 'employment outlay ratio' of the  $i$ -th sector. Then we obtain

$$\frac{\sum_j a'_{oj} [X_j(t+1) - X_j(t)]}{\sum_j X_j(t)} = \alpha(t) \sum_i \gamma_i \mu_i(t),$$



# SOME OBSERVATIONS ON INPUT-OUTPUT ANALYSIS

or, by introducing the average employment-outlay ratio of the national economy

$$\gamma(t) = \sum_i \gamma_i \mu_i(t), \quad \dots (7.18)$$

$$\frac{\sum_j a'_{0j} [X_j(t+1) - X_j(t)]}{\sum_j X_j(t)} = \alpha(t) \gamma(t). \quad \dots (7.19)$$

The left hand side of (7.19) indicates the increment of total employment from one unit period to the next in relation to the value of the gross national product in the initial unit period. Let us write,

$$a'_0(t) = \frac{\sum_j a'_{0j} X_j(t)}{\sum_j X_j(t)}, \quad \dots (7.20)$$

i.e., the average employment coefficient of the national economy. Substituting this into (7.19) we obtain the rate of increase of total employment from one unit period to the next;

$$\frac{\sum_j a'_{0j} X_j(t+1) - X_j(t)}{\sum_j a'_{0j} X_j(t)} = \frac{\alpha(t) \gamma(t)}{a'_0(t)},$$

or, denoting the left hand side by  $\rho(t)$ ,

$$\rho(t) = \frac{\alpha(t) \gamma(t)}{a'_0(t)}. \quad \dots (7.21)$$

Thus we find that the rate of increase of total employment is the product of the rate of investment and the average employment-outlay ratio divided by the average employment coefficient of the national economy.

The total employment in unit period  $t$  is related to the total employment in the initial unit period  $t_0$  ( $t > t_0$ ) by the formula

$$\sum_j a'_{0j} X_j(t) = \prod_{t=t_0}^t \left[ 1 + \frac{\alpha(t) \gamma(t)}{a'_0(t)} \right] \sum_j a'_{0j} X_j(t_0). \quad \dots (7.22)$$

Comparing (7.21) with (7.9), we can establish a relation between the rate of increase of employment and the rate of increase of national income (or, which is the same, of gross national product.) Denote by  $v(t)$  the ratio of these two rates, i.e.,

$$v(t) = \frac{\rho(t)}{r(t)}; \quad \dots (7.23)$$

we have

$$v(t) = \frac{1}{a'(t)} \cdot \frac{\gamma(t)}{\beta(t)}; \quad \dots (7.24)$$

i.e., this ratio is proportional to the ratio of the average employment-outlay ratio and the average output-outlay ratio.

Total employment grows faster, equal or slower than national income according as to whether

$$\frac{\gamma(t)}{a'_0(t)} \begin{matrix} > \\ = \\ < \end{matrix} \beta(t). \quad \dots (7.25)$$

However,  $\gamma(t)$  and  $\beta(t)$  are averages depending on the allocation of the total investment outlay among the various sectors of the national economy. Remembering (7.8) and (7.18) we have

$$\nu(t) = \frac{1}{a'_0(t)} \frac{\sum_i \gamma_i \mu_i(t)}{\sum_i \beta_i \mu_i(t)} \quad \dots (2.26)$$

Since the coefficients  $\gamma_i$  and  $\beta_i$  are determined by technological conditions and  $a'_0(t)$  is determined by the employment coefficients  $a'_{0j}$  and by the way the national product is composed of outputs of the various sectors,  $\nu(t)$  can be influenced only by a proper choice of the allocation of investment fractions  $\mu_i(t)$ .

In order to obtain the greatest rate of increase of national income (or of gross-national output) the allocation fractions  $\mu_i(t)$  have to be chosen so as to maximize the average overall output-outlay ratio  $\beta(t)$ . In order to achieve this, investment outlays must be allocated to the sectors with the highest overall out-lay ratios,  $\beta_i$ .

In order to obtain the greatest possible rate of increase of total employment the allocation fraction  $\mu_i(t)$  have to be chosen so as to maximize the average employment outlay ratio  $\gamma(t)$ . This requires that the investment outlays be allocated to the sectors with the highest overall employment outlay ratio  $\gamma_i$ .

These considerations refer to the rate of increase of national income or of total employment in a given unit period  $t$ . If the goal of the policy is to obtain the greatest possible increase of total employment after a longer period of time an additional factor has to be brought into consideration. From (7.21) we see that the rate of increase in total employment is proportional to  $\alpha(t)$  i.e., the rate of investment in the unit period. The rate of investment, however, may depend on the national income, because an increase in national income makes it possible to have a greater rate of investment.

Consequently, it may be possible to obtain in the long run a greater increase in total employment by allocating investment outlays not in a way which produces immediately the greatest rate of growth of total employment but in a way which produces the greatest rate of increase of national income. The slower rate of increase of employment in the beginning period is then over-compensated by a more rapid rate of increase of employment in the later period due to an increased rate of investment.

For instance, let

$$\alpha(t) = cI(t), \quad \dots (7.27)$$

# SOME OBSERVATIONS ON INPUT-OUTPUT ANALYSIS

where  $I(t) = \sum x_j(t)$  is the national income in unit period  $t$  and  $c$  is a factor of proportionality ( $0 < c < 1$ ). Then,

$$\rho(t) = \frac{cI(t)\gamma(t)}{a'_0(t)}. \quad \dots (7.28)$$

Taking into account relation (7.14), we find that in any given unit period  $t_k$  ( $t_k < t_0$ ) the rate of increase of total employment is

$$\rho(t_k) = c \frac{\gamma(t_k)}{a'_0(t_k)} I(t_0) \prod_{t=t_0}^{t_k} (1+r(t)), \quad \dots (7.29)$$

where  $I(t_0)$  is the national income in the initial unit period,  $t_0$ .

Thus the rate of increase of total employment in any given unit period is proportional to the increase of national income which took place between the initial unit period and the unit period under consideration.

In expression (7.29)  $\gamma(t_k)$  depends on the values of the investment allocation fractions  $\mu_i(t_k)$  ( $i = 1, \dots, n$ ) in unit period  $t_k$  whereas  $r(t)$  depends on the values of the allocation of investment fractions  $\mu_i(t)$  in all the unit periods from  $t_0$  to  $t_k$ . This can be seen immediately from the formulae (7.8), (7.9), and (7.18). A change of the values of the allocation (of investment) fractions in each period from  $t_0$  to  $t_k$  thus produces a change in the rate of increase of total employment in unit period  $t_k$  equal to

$$d\rho(t_k) = \frac{c}{a'_0(t_k)} I(t_0) \left[ \prod_{t=t_0}^{t_k} (1+r(t)) d\gamma(t_k) + \gamma(t_k) d \prod_{t=t_0}^{t_k} (1+r(t)) \right]. \quad \dots (7.30)$$

The change is positive zero or negative according to the sign of the expression in braces on the right hand side, i.e., according as to whether

$$\frac{d \prod_{t=t_0}^{t_k} (1+r(t))}{\prod_{t=t_0}^{t_k} (1+r(t))} \begin{matrix} > \\ = \\ < \end{matrix} - \frac{d\gamma(t_k)}{\gamma(t_k)}. \quad \dots (7.31)$$

The left hand side of (7.31) can be written in the form

$$d \log \prod_{t=t_0}^{t_k} (1+r(t)) = \sum_{t=t_0}^{t_k} \frac{dr(t)}{1+r(t)}.$$

Hence, the expression (7.31) becomes

$$\sum_{t=t_0}^{t_k} \frac{dr(t)}{1+r(t)} \begin{matrix} > \\ = \\ < \end{matrix} - \frac{d\gamma(t_k)}{\gamma(t_k)}. \quad \dots (7.32)$$



Let us start with values of the allocation of investment fractions which in each unit period from  $t_0$  to  $t_k$  maximize the average employment-outlay ratio  $\gamma(t)$ . Then change these fractions so as to maximize  $r(t)$ . In each unit period  $dr(t) > 0$  and  $d\gamma(t_k) < 0$  (except in the trivial case when  $\gamma(t) = \beta(t)$  in each unit period, in which case  $dr(t) = 0 = d\gamma(t)$ ). Thus the left hand side of (7.32) increases monotonously with the value of  $t_k$ . By choosing  $t_k$  large enough it is possible to make the left hand side in (7.32) greater than the right hand side, i.e., to achieve a greater rate of increase of total employment than would be the case if the investment allocation fractions were chosen so as to maximize in each unit period the immediate effect on total employment.

Total employment in the unit period  $t_s (t_s \geq t_k \geq t_0)$  is according to (7.22)

$$\sum_j a'_{oj} X_j(t_s) = \prod_{t_k=t_0}^{t_s} [1 + \rho(t_k)] \sum_j a'_{oj} X_j(t_0). \quad \dots (7.33)$$

Taking logarithms, we find

$$d \log \sum_j a'_{oj} X_j(t_s) = \sum_{t_k=t_0}^{t_s} \frac{d\rho(t_k)}{1 + \rho(t_k)} + \text{constant}. \quad \dots (7.34)$$

As we have seen, a change of the allocation of investment fractions designed to maximize  $r(t)$  in each unit period leads to  $d\rho(t_k) > 0$  from a certain unit period on wards. Beginning with that unit period the right hand side of (7.34) increases monotonously, with the value of  $t_s$ . By choosing  $t_s$  large enough it is possible to make (7.34) positive, i.e., to make total employment larger than would be the case if the rate of increase of national income were not maximised in each unit period.

Denote by  $t_c$  the critical value of  $t_s$  at which the expression starts becoming positive. Over planning periods which are shorter than  $t_c - t_0$  the greatest possible total employment is obtained by allocating investment outlays among the various sectors of the national economy so as to maximise in each unit period  $\gamma(t)$  by directing them always to the sectors with greatest employment-outlay ratios. Over planning periods exceeding  $t_c - t_0$  the greatest possible total employment is obtained by maximising in each unit period  $r(t)$ , i.e., by allocating investment outlays always to the sectors with the greatest output-outlay ratios.

More complicated conditions for allocation of investment outlays among the various sectors of the national economy are obtained when the principal goal of the policy i.e., greatest possible increase of national income or of total employment during a period of time, is subject to additional conditions imposed like, for instance, a certain predetermined rate of growth of consumption. Such problems can be solved on the basis of the relations established in this chapter by means of the techniques of linear programming.

# THE USE OF A SHORT-TERM ECONOMETRIC MODEL FOR INDIAN ECONOMIC POLICY

By J. TINBERGEN

*Planning Bureau of the Netherlands, The Hague*

## 1. THE SECOND FIVE-YEAR PLAN AND SHORT-TERM ADAPTATIONS

Most of the figures and recommendations on the Second Five Year Plan so far published refer to its targets. In the Draft Outline, figures can be found about the increase in national income and employment it is proposed to strive for, the investments that will have to be made in order to create the necessary productive capacity and the financing of that investment, particularly in the public sector. Recommendations are made with respect to a number of qualitative subjects such as education, training, social services and so on. Not all of the figures and recommendations refer to targets; but many of them do. Much less is to be found on the instruments to be used. To be sure, public investments may be called instruments in this respect. Their realization will depend on the financial means available and it is well-known that their enumeration in the Draft Outline is not yet complete; also the true instruments here are the tax rates rather than the total tax receipts<sup>1</sup> and these rates have not been specified yet. Private investments, to quote another example, cannot be called instruments either; they are themselves the consequences of a number of circumstances, among which the instruments of economic policy. Very probably many non-specified instruments in the field of restrictions and allocations are also going to play a role in the execution of this Five Year Plan and these instruments have not yet been specified.

They could hardly be specified in much detail, since their function is essentially to bring about the short-term adaptations needed in every economy, whether in development or not. Every economy is subject to a number of almost unforeseeable short-term disturbances: changes in crops, in foreign commercial policy, in consumer and investor attitudes, etc. It may happen, in a certain year, that consumption of the population exceeds the level desired by the Five Year Plan leaving too little for investment. In order to warrant the Plan's execution the government will have to make short-term adaptations in its taxation and spending programme. It may happen that certain expenditures have been underestimated and that additional finance is needed. The same problem then arises. It may be that the balance of payments shows an undesirable development: imports may have to be restricted or exports to be stimulated. For these short-term adaptations to be performed

---

<sup>1</sup> See B. Hansen, *Finanspolitikens ekonomiska teori*, Stockholm 1955, p. 49 ff.

in an orderly way and on time, a systematic preparation may be useful. It is the objective of this paper to give a brief survey of the methods used in this field in the Netherlands and to give a discussion of the question whether these methods may also be adopted by a country so different as India.

## 2. SHORT-TERM MODELS USED IN POLICY DESIGN

In the Netherlands, where the Second World War had caused considerable damage, and where several structural changes had to be brought about, economic policy since 1945 has been prepared in a more systematic way than before. As in most West European countries, annual calculations of national income, and its origin and target, have been made together with social accounts. Some further steps have, however, been made also. Along with statistical estimates for the immediate past, estimates for future years or budgets have also been made for the nation as a whole. In addition, estimates have been made for such future years of the consequences of certain measures of economic policy. These latter calculations incidentally also were made in other countries; in Holland they were, however, based on econometric models.<sup>1</sup>

To be sure these models are still very rough presentations of the most important relations only between the most important variables of the economy. Their use has been questioned by several critics and it should be stated that they are meant to help in political decisions, and are not to be used as a substitute for common sense. It seems to the author that their advantages lie in their being more explicit than most other instruments of analysis, and that they permit, once they have been constructed, the execution of a large number of alternative calculations. Such alternatives are useful in a double way. They enable the policy-maker to study the consequences of variations in the policy programme and, at the same time, they are illustrative of the consequences of errors of estimation, that is, of the range of uncertainty in the results.

Models are more explicit in that they force their authors to state the problem of policy they are going to discuss among other things by an enumeration of all the variables considered and of the relations assumed to exist between them.<sup>2</sup> In policy problems they also force their authors to state explicitly the targets and the instruments of the policies considered. In view of the multitude of conceivable problems in economic policy and of the confusion caused by an incomplete setting of the problems considered, such explicit statements are not only very useful but almost indispensable.

---

<sup>1</sup> See Central Planning Bureau, *Central Economic Plan 1955*, The Hague, 1955.

<sup>2</sup> For examples see L. R. Klein and A. S. Goldberger, *An Econometric Model of the United States*, Amsterdam, 1955.



## SHORT-TERM ECONOMETRIC MODEL FOR INDIAN ECONOMIC POLICY

### 3. THE CONSTRUCTION OF SHORT-TERM ECONOMETRIC MODELS

It cannot be the purpose of a brief article to give a detailed account of the construction of the models just mentioned. These details may be found in the existing literature.<sup>1</sup> A summary with a few practical hints only may be given here. The construction essentially consists of a theoretical and a statistical part, which are intimately interwoven: in this section the theoretical part will be described, but it should not be overlooked that each theoretical contribution has to be tested statistically and if rejected by the test, to be amended, until a model is found which satisfies the standards set both from the theoretical and from the statistical point of view.

The first instalment is a choice of the variables which are considered important for the purpose of the model construction. This choice has to depend on the type of policy to be considered and anyhow the target and instrument variables have to enter into the list of variables. To the variables we are interested in, have to be added the variables which are assumed to be relevant for the explanation of the fluctuations of the target variables. These may be variables we are not interested in a priori. Suppose national income is one of our target variables. National income will depend on the volume of investment (in the Keynesian way), and we are interested in investment since we consider it as one of the instruments; so we will have that in our list. But apart from investment, national income may also depend on exports which for their own sake we are not interested in so much. They have to be added to the list, since they are relevant to income. Again we may have consumption in our list since consumption is one of the ultimate ends of our policy; consumption will, however, depend on the price level, and so the price level has to be taken into consideration also. Which variables should finally be included in the list, partly depends on our statistical testing of the relations we assume. Variables, which a priori may seem to be important, may turn out to be otherwise (e.g., the interest rate); on the other hand, variables which we may not consider important, may, on closer examination, turn out to be quite relevant.

### 4. STATISTICAL ESTIMATION OF CONSTANTS

It is the essence of statistical estimation that the fluctuations in some of the variables are explained with the aid of certain constants and by the fluctuations in other variables. Such constants may either take the form of initial values of some variables, e.g., the initial value of national income and national capital and of the price level etc., or the form of so-called coefficients, e.g., the propensity to consume, the import quota of certain industries etc. These constants, which often are only approximately constant, or dependent on other, more hidden, constants with variable "weights", represent the structure of the economy under investigation. Their statistical estimation is one of the most important steps in the whole procedure.

---

<sup>1</sup> See Klein and Goldberger, loc. cit.; or J. Tinbergen, *Business Cycles in the United Kingdom, 1870-1910*, Amsterdam 1951.

There are several methods of estimation ranging from simple ones to very complicated ones. It is one of the objectives of the work under review to decide which method to use. It would be a mistake to believe that only time series analysis and in particular the most modern methods are used. Certainly, time series analysis is an important method. In addition, however "structural analysis" may be helpful i.e., statistical analysis of the composition of certain totals or the differences in behaviour of groups existing at the same time. So, the now famous capital coefficient (or  $\frac{1}{\beta}$ ) may be estimated with the aid of very simple methods. Or the propensity to consume may be derived from family budget statistics. Or, to quote a rather important source of knowledge, international comparison may be used.<sup>1</sup>

In certain cases a still more direct estimation of coefficients is possible, e.g., for tax rates. The average marginal rate of direct taxation, to quote an example, may be derived from the tax regulations themselves.

In the absence of any solid statistical basis, one possibility to obtain numerical estimates of certain coefficients may be to ask a number of experts in the matter for a personal guess. This illustrates that in practical work, inventiveness and imagination will have to play their part.

##### 5. FLEXIBILITY OF THE METHOD : APPLICATION TO INDIA

It is often believed that models built for one type of country cannot be of much use for countries of a very different type. Such a statement may cause confusion. It is true of course if one would think of applying the equations of exactly the same numerical form found for one country to another country; that is of no use. The statement need not be true in another version, namely if only the mathematical shape of the equations is maintained, but other numerical values of the coefficients are applied. Already in this way important differences in structure can be properly taken account of. Some of the most important structural constants used in the Netherlands model are the following.

- (i) The marginal import content of national product  $\mu$  :
- (ii) The portion  $\Lambda$  of national income accruing to employees:
- (iii) Government expenditure  $G^0$  in the initial period (may be expressed as a percentage of national product at market prices);
- (iv) The marginal rate of spending  $1-\sigma$ .

By a variation in these constants we may obtain formulae for a country showing a completely different import content, a completely different distribution of national income between employees and independents, a completely different size of the government sector in comparison to the private sector and a completely

<sup>1</sup> Colin Clark was one of the first to apply this method on a large scale (see his *Economics of 1960*, London 1942). A recent example is L. Jureen: "Long-Term Trends in Food Consumption: A Multi-Country Study", *Econometrica* 24, (1956) p.1.



different rate of spending. All of these are already important characteristics of an economy, may be even the most important ones. In fact, if one were to enumerate the systematic differences between the Netherlands and India one would probably find these characteristics to be of importance. But the fact of their being widely different for the two countries does not mean that the method of econometric models cannot be applied to India. In this respect the method possesses every flexibility that could be desired. Several other examples could be added, such as the different composition of imports and to some extent of exports. They will not cause any difficulty in the application of the Dutch model to India.

There are, of course, other difficulties of a more fundamental character. These we will now discuss.

a) First of all, one may wonder whether or not India, being such a large country, is *far less homogeneous* and whether this would not frustrate the method to some extent. Upon some further reflexion it is clear that no general pronouncement can be made.

It is certainly probable that the price level or the wage level will be much less uniform in India than in a small country like the Netherlands. This may imply that averages are less meaningful and hence statistical relations are less clear. But it need not imply this consequence, since the very fact of the size may permit the "application of the law of large numbers" in India rather than in Netherlands. In fact it appears that the fluctuations in wages and those in prices show a very satisfactory correlation (cf. section 6). Trying it out seems to be the conclusion and this is the very essence of the econometric method.

In other respects one may even doubt whether there is at all more heterogeneity in India than in the Netherlands. The number of industries, e.g., may well be higher in the latter country than in the former, where 70% of the population is working in agriculture.

b) *The preponderant influence of agricultural production* may entail certain difficulties. One example is the fluctuations in quality of the crops, a phenomenon less pronounced in other industries. It may well be that the fluctuations in prices may to some extent be ascribed to these fluctuations, and since hardly any data on qualities seem to exist, this may endanger the application of our methods. But then one should not forget that so far the best results of econometric research have been obtained in the agricultural sector and mainly for another reason; in this sector random fluctuations are much larger than elsewhere and this reduces intercorrelations between explanatory variables. The separation of supply and demand factors, e.g., of the price of a good like cotton has been very successful for this very reason.

c) In a general way it may be assumed that the behaviour of the Indian population in economic matters is *less rational* than that of an industrialized population like the Dutch. As a consequence, is the econometric method less reliable? It again seems to depend. To the extent that, for this reason, random deviations



in demand pattern would be larger, it may indeed be so. But there are two counter-acting factors. To the extent that instead of rationality there is traditional behaviour, this may even mean more constancy in behaviour, i.e., low elasticity of demand. For another reason also, there may be less divergency between various families or even regions, namely, greater poverty. It is well-known that poorer people have less possibilities to choose their diet since the bare necessities have to be satisfied first. The conclusion, therefore, cannot be but : the proof of the (econometric) pudding is in the eating.

d) The existence of a *barter sector* much larger than in many other economies seems to be a further point to consider. This seems to reduce the interdependency between the economic variables, since in a way it is equivalent to the existence of a larger number of unconnected (or hardly connected) units. This is, however, only one way of putting it. In a barter economy, more than elsewhere, there is the absence of credit facilities and, therefore, perhaps a more stringent tie between income and expenditure. As far as this relation is concerned, therefore, nothing much is to be feared.

e) A real difficulty seems to be the more pronounced influence of *speculative trading* in bulk commodities. It is one of the characteristics of an underdeveloped country that a larger part of entrepreneurs are hunting for short-term profits instead of steady, moderate profits after a longer gestation period. It is also well-known that in certain circumstances speculation has been a factor of considerable influence in price formation. And it should be recognised that so far econometric analysis of speculative buying has had very moderate success only<sup>1</sup>. This then is a weak spot. The conclusion might be that more intensive research into this phenomenon is needed—or that the interests of econometrists and of the country at large to eliminate speculation are parallel !

f) A final objection to applying econometric methods to a country like India might seem to be the *quality of statistical data*. Here again careful distinction is necessary. Certain basic data may be better in India than elsewhere. The well-known example is crop statistics. Still there are other data which very probably are of poorer quality in India than say in Europe, e.g., accounting data for small-scale industries, trade and agriculture. India is facing this problem and no doubt the quality will gradually improve. What are the consequences of bad data ? Evidently the amplitude of certain types of errors will be larger, leading to less reliable estimates. It cannot be denied that this is a disadvantage; the general conclusion should be the same as with (a), (b) and (c) : try it out. A wrong idea rather frequent with laymen should be mentioned. A larger amplitude of the error component in a correlation does not imply that there is a systematic bias in the consistent estimate of a coefficient. It depends on other circumstances whether such a bias will occur. Under

---

<sup>1</sup> cf. H. Rijken van Olst, dissertation Rotterdam

## SHORT-TERM ECONOMETRIC MODEL FOR INDIAN ECONOMIC POLICY

certain conditions the policy conclusion will not be influenced by the presence of stochastic deviations.<sup>1</sup>

Our general conclusion from the brief analysis of the differences between India and the Netherlands (or other similar countries) seems to be that no systematic bias, which may be of disadvantage to the application of econometric methods to India, seems to exist. There are certain disadvantages to be expected, but there are also counteracting factors and it seems difficult to forecast the final result. The method has to be tried out.

### 6. SOME EXAMPLES OF RELATIONS

The construction of a complete system of econometric relations describing at least with some accuracy the dynamics of short-term economic fluctuations is not a one-man job: it can only be done by teamwork and hence by an institute. The Indian Statistical Institute would seem to be the most appropriate place for such an enterprise, with the co-operation of the Reserve Bank of India, where interesting work has already been done,<sup>2</sup> and the co-operation of agencies with specialized knowledge of certain sectors of the economy. Still it seems that pioneering attempts may well be made by individuals — the works just quoted being examples. I am happy to report that some pioneering work has been done by Mr. Narasimham in his thesis.<sup>3</sup> It is with his permission that I reproduce some of his results in order to illustrate the type of analysis discussed. The first relation is the one already mentioned in section 5 (a), a relation meant to "explain" the short-term fluctuations in the wage level. It has been assumed, from the experience gained in other countries, that the most important explanatory variables are cost of living and employment, whereas the slowly changing factors are represented by a trend component. Upon this assumption a fairly good explanation of wage fluctuations can be found and the correlation is highest if we take the reaction coefficient of prices to be 0.23, i.e., a 1% increase in prices leads to a 0.23% increase in wage rates. The influence of employment fluctuations on the wage rate is expressed by a flexibility coefficient of 0.84, if our theory is correct and there is a downward trend.

This trend implies that real wages have fallen on the average over the period 1923–1950, a phenomenon not found in Western countries and probably related to the population pressure in India.

As a second example, let me quote one of the technical equations in Mr. Narasimham's book namely, the relation assumed between the volume of employment  $a_t$  and the volumes of production of consumer goods ( $u_t''$ ) and investment goods

---

<sup>1</sup> G. H. Theil: "Econometric Models and Welfare Maximization", *Weltw Archiv*, 72 (1954) p.60.

<sup>2</sup> Especially by Messrs. Sastry and Murti.

<sup>3</sup> N. V. A. Narasimham, *A Short Term Planning Model for India*, North Holland Publishing Company Amsterdam 1956.



( $v_t''$ ). All the three variables are expressed in crores of constant rupees (of 1938) and the relation reads :

$$a_t = 0.026a_t'' + 0.20v_{t-\frac{1}{2}}'' + 0.025$$

The meaning of this formula is that variations in the production of consumer goods cause parallel variations in employment of about 2.6% only, whereas variations in the production of equipment goods cause variation in employment of about 20% and with a lag of half year. The figures 2.6% and 20% may be considered the marginal labour content of consumer and investment goods, respectively. The figure for investment goods seems to be of normal order of magnitude; the very low level of the figure for consumer goods is probably a reflection of the large number of enterprises working without hired labour in a country like India. Even the figure of 20% for investment goods industries is rather low in comparison to developed countries but may be explained by a relatively high import content in this industry and perhaps to some extent also by the existence of a large number of small enterprises.

In certain equations the variables to be introduced may be different from those used in the Netherlands system. One example is the occurrence of crop yields in the equation "explaining" price levels of home products or exports. For a country where crop production is so preponderant, this is self-evident. It may not be Indian crops only, but the crops of competing countries as well that have to be taken account of. Cotton prices, for example, may be influenced by American, Brazilian and Egyptian crops and probably to a different degree.

Other variables may also occur because of the use of a greater number of the instruments of economic policy in India than in the Netherlands. As a rule, the more pressing the problems of a country the more instruments will be used. This is why periods of war are characterized by the use of more instruments, especially of the type of quantitative restrictions, than other periods. Even if certain instruments cannot be used in India, there is every chance that such instrument will have to be introduced at least tentatively into the planning work in order to study the probable effects. Thus one might imagine the introduction of further import restrictions as means of policy, leading to an "import equation" expressing nothing but the "rations" imposed on the various sectors of the economy. Such rations will, however, have their repercussions on the expenditures made by these sectors in the home market, and the effects of these repercussions will have to be expressed in some of the other equations representing these expenditures. It is no use working this out in any detail before a choice has been made as to the instruments to be introduced into the system.

*Paper received: August, 1956.*



# MISCELLANEOUS

## APPROXIMATE DISTRIBUTION OF CERTAIN LINEAR FUNCTION OF ORDER STATISTICS<sup>1</sup>

By K. C. SEAL

*University of North Carolina, Chapel Hill*

### 1. INTRODUCTION

Suppose  $Y_{(1)} < \dots < Y_{(n)}$  are  $n$  order statistics from a standard normal population  $N(0, 1)$  and that  $Y_0$  is an independent random variable obeying  $N(0, 1)$ . Corresponding to each  $n$  suppose there is given a set of  $n$  constants  $(c_{1n}, c_{2n}, \dots, c_{nn})$  such that  $c_{in} \geq 0$  and  $\sum_{i=1}^n c_{in} = 1$ . The purpose of this paper is to show that

$$\sum_{i=1}^n c_{in} Y_{(i)} - Y_0 - E \left( \sum_{i=1}^n c_{in} Y_{(i)} \right)$$

has an asymptotic normal distribution and from some further results derived below, it appears that this distribution is approximately normal for all  $n$ .

### 2. ASYMPTOTIC NORMALITY

We show at first that the limiting distribution of

$$\sum_{i=1}^n a_{in} Y_{(i)} + Y_0 - E \left( \sum_{i=1}^n a_{in} Y_{(i)} \right),$$

where  $\sum_{i=1}^n |a_{in}| \leq M$ , an arbitrary constant, is  $N(0, 1)$ .

We easily get

$$\text{Var} \left( \sum_{i=1}^n a_{in} Y_{(i)} \right) \dots (2.1)$$

$$= \sum_{i=1}^n a_{in}^2 \text{Var} (Y_{(i)}) + \sum_{i=1}^n \sum_{j=1, j \neq i}^n a_{in} a_{jn} \text{cov} (Y_{(i)}, Y_{(j)}).$$

---

<sup>1</sup> This research was supported in part by the United States Air Force through the Office of Scientific Research of the Air Research and Development Command.

From Cramér (1946) pp. 374-377, it can be readily verified that for  $n$  sufficiently large

$$\text{Var } (Y_{(n)}) \geq \text{Var } (Y_{(i)}), \quad i = 1, 2, \dots, n-1.$$

Hence by (2.1) it follows that

$$\text{Var} \left( \sum_{i=1}^n a_{in} Y_{(i)} \right) \leq \left( \sum_{i=1}^n |a_{in}| \right)^2 \text{Var } (Y_{(n)}) \leq \frac{M^2}{k \log n},$$

where  $k$  is some constant. Hence  $\lim_{n \rightarrow \infty} \text{Var} \left( \sum_{i=1}^n a_{in} Y_{(i)} \right) = 0$  and so by Cramér (1946),

pp. 253-55, it follows that the limiting distribution of  $\sum_{i=1}^n a_{in} Y_{(i)} + Y_0 - E \left( \sum_{i=1}^n a_{in} Y_{(i)} \right)$  will be the same as that of  $Y_0$ , thus being  $N(0, 1)$ .

From this result it immediately follows that the limiting distribution of

$$\sum_{i=1}^n c_{in} Y_{(i)} - Y_0 - E \left( \sum_{i=1}^n c_{in} Y_{(i)} \right),$$

where

$$c_{in} \geq 0, \quad i = 1, \dots, n \quad \text{and} \quad \sum_{i=1}^n c_{in} = 1, \quad \text{is } N(0, 1).$$

### 3. APPROXIMATE DISTRIBUTION.

We shall now show that the distribution of  $Y_{(n)} - Y_0 - E(Y_{(n)})$  can be taken to be very nearly normal for all  $n$ . Let us denote by  $\mu_{r/n}$  and  $\mu_{r(n)}$  the  $r$ -th corrected moment of  $Y_{(n)}$  and  $Y_{(n)} - Y_0$  respectively. Also  $\beta_1$  and  $\beta_2$  coefficients of  $Y_{(n)}$  and  $Y_{(n)} - Y_0$  will be written as  $\beta_{1/n}$  and  $\beta_{1(n)}$ ,  $i = 1, 2$ , respectively. Clearly the expected values of  $Y_{(n)}$  and  $Y_{(n)} - Y_0$  are identical. It can be easily verified that

$$\begin{aligned} \mu_{2(n)} &= 1 + \mu_{2/n}, \\ \mu_{3(n)} &= \mu_{3/n}, \\ \mu_{4(n)} &= \mu_{4/n} + 6\mu_{2/n} + 3. \end{aligned} \quad \dots \quad (3.1)$$

Let

$$\gamma_{1(n)} = \sqrt{\beta_{1(n)}}, \quad \gamma_{1/n} = \sqrt{\beta_{1/n}};$$

and

$$\gamma_{2(n)} = \beta_{2(n)} - 3, \quad \gamma_{2/n} = \beta_{2/n} - 3.$$

By (3.1) we obtain after a little simplification

$$\gamma_{1(n)} = \gamma_{1/n} \cdot \left( \frac{\mu_{2/n}}{1 + \mu_{2/n}} \right)^{3/2}, \quad \dots \quad (3.2)$$

$$\gamma_{2(n)} = \gamma_{2/n} \cdot \left( \frac{\mu_{2/n}}{1 + \mu_{2/n}} \right)^3.$$

# LINEAR FUNCTION OF ORDER STATISTICS

Tippett (1925) has given values of  $\beta_{1/n}$  and  $\beta_{2/n}$  for  $n$  ranging between 2 and 1000. Hence by (3.2) the following table can be constructed.

TABLE 3.1 VALUES OF  $\beta_{1(n)}$  AND  $\beta_{2(n)}$

$n$	$\beta_{1(n)}$	$\beta_{2(n)}$
2	0.00127	3.010
5	0.00272	3.019
10	0.00282	3.022
20	0.00253	3.022
60	0.00189	3.020
100	0.00162	3.019
200	0.00131	3.017
500	0.00100	3.015
1000	0.00082	3.013

The above values of  $\beta_{1(n)}$  and  $\beta_{2(n)}$  in Table 3.1 suggests that the distribution of  $Y_{(n)} - Y_0$  for  $2 \leq n \leq 1000$  is approximately normal. We have seen in Section 2 that the distribution of  $Y_{(n)} - Y_0 - E(Y_{(n)})$  is asymptotically normal; also it is clear that for  $n = 1$  the distribution is exactly normal. From Table 3.1 it may be noticed that the greatest departure from normality occurs between  $n = 5$  and  $n = 20$ . Thus, although the distribution of  $Y_{(n)}$  departs more and more from normality with increasing  $n$ , the distribution of  $Y_{(n)} - Y_0$  comes closer to normality beyond  $n = 20$ .

From the results of Gumbel (1936) it is seen that for  $N(0, 1)$  the skewness and kurtosis of order statistics increase, at least for very large samples, as we move further away from the median. It is presumed that this result is true for any sample size but this could not be demonstrated by any simple method. Further it appears that, for every  $n$ , the greatest departure from normality among all linear functions  $\sum_{i=1}^n c_{in} Y_{(i)}$ , where  $c_{in} \geq 0$ ,  $i = 1, \dots, n$

and  $\sum_{i=1}^n c_{in} = 1$ , will happen for  $Y_{(n)}$ . Assuming this conjecture to be true it will then follow from the above results that any linear function

$$\sum_{i=1}^n c_{in} Y_{(i)} - Y_0 - E\left(\sum_{i=1}^n c_{in} Y_{(i)}\right),$$

with similar restrictions on  $c_{in}$ 's, will have approximately a normal distribution - the order of approximation to normality will not be cruder than that of  $Y_{(n)} - Y_0 - E(Y_{(n)})$ . To verify this result experimentally, the following model sampling experiment was performed.

A thousand random samples of size 4 were drawn from  $N(0, 1)$  with the help of tables provided by Mahalanobis, Bose, Ray and Banerji (1934). Corresponding to each sample another independently drawn observation of  $Y_0$  was also taken. Four observations were ranked in an increasing order of magnitude and frequency distribution of the following six statistics were evaluated:

$$(a) \quad 0.25 [Y_{(1)} + Y_{(2)} + Y_{(3)} + Y_{(4)}] - Y_0,$$

$$(b) \quad 0.25 Y_{(2)} + 0.25 Y_{(3)} + 0.50 Y_{(4)} - Y_0,$$

$$(c) \quad 0.25 Y_{(3)} + 0.75 Y_{(4)} - Y_0,$$



- (d)  $Y_{(4)} - Y_0$ ,  
 (e)  $0.50 Y_{(3)} + 0.50 Y_{(4)} - Y_0$ ,  
 (f)  $0.50 Y_{(1)} + 0.50 Y_{(4)} - Y_0$ .

The  $\beta_1$  and  $\beta_2$  coefficients for the above six distributions are given in Table 3.2.

TABLE 3.2.  $\beta_1$  AND  $\beta_2$  OF SIX CHOSEN STATISTICS

linear function	$\beta_1$	$\beta_2$	$\chi^2$	df
(a) $0.25[Y_{(1)} + Y_{(2)} + Y_{(3)} + Y_{(4)}] - Y_0$	0.011	3.067	25.73	25
(b) $0.25 Y_{(2)} + 0.25 Y_{(3)} + 0.50 Y_{(4)} - Y_0$	0.00005	3.116	28.03	26
(c) $0.25 Y_{(3)} + 0.75 Y_{(4)} - Y_0$	0.00202	3.101	23.89	26
(d) $Y_{(4)} - Y_0$	0.0318	3.285	24.78	27
(e) $0.50 Y_{(3)} + 0.50 Y_{(4)} - Y_0$	0.00612	3.465	25.58	25
(f) $0.50 Y_{(1)} + 0.50 Y_{(4)} - Y_0$	0.0200	3.114	30.49	25

The exact values of  $\beta_1$  and  $\beta_2$  for the frequency function of (a) are 0 and 3 respectively and for that of (d), Table 3.1 suggests that  $\beta_1$  lies between 0.00127 and 0.00272 and that  $\beta_2$  lies between 3.010 and 3.019. Hence comparing these known values with the corresponding values of Table 3.2 an idea of the magnitude of sampling fluctuation is obtained. Another method for testing this approximate normality will be to fit an appropriate normal distribution to the observed frequency distribution of each of these six statistics and test the goodness of fit by the usual  $\chi^2$ . Although it is definitely known that distribution of none of these statistics excepting the first is exactly normal, the above  $\chi^2$  test may be applied on the assumption that non-significance of  $\chi^2$  values for testing goodness of fit only implies that the null hypothesis is approximately true. The  $\chi^2$  values along with their respective degrees of freedom thus obtained are shown in columns 4 and 5 of Table 3.2. It is seen that the probability levels are all greater than .30 excepting that of the last linear function (f) where it is nearly .20. Thus each of these  $\chi^2$  is insignificant. It therefore appears that our conjecture about approximate normality of

$$\sum_{i=1}^n c_{in} Y_{(i)} - Y_0 - E \left( \sum_{i=1}^n c_{in} Y_{(i)} \right)$$

may be correct for  $n = 4$ . Detailed study about this approximate normality has been made by the author (Seal (1954)) and this property of approximate normality was used in a decision problem.

I wish to thank Professor R. C. Bose for his kind help and advice in the course of my investigation.

#### REFERENCES

- CRAMÉR, H. (1946): *Mathematical Methods of Statistics*, Princeton University Press.  
 GUMBEL, E. J. (1936): Les Valeurs extremes des distributions statistiques. *Ann. Inst. Henri Poincaré*, 5, 115.  
 MAHALANOBIS, P. C., BOSE, S. S., RAY, P. R. and BANERJI, S. K. (1934): Tables of random samples from a normal population. *Sankhyā*, 1, 289.  
 SEAL, K. C. (1954): On a class of decision procedures for ranking means, *Unpublished Ph. D. Thesis*. University of North Carolina, Chapel Hill, U.S.A..  
 TIPPETT, L. H. C. (1925): The extreme individuals and the range of samples taken from a normal Population. *Biometrika*, 17, 364.

Paper received: April, 1956.

# ON THE UNBOUNDEDNESS OF INFINITELY DIVISIBLE LAWS

By S. D. CHATTERJEE AND R. P. PAKSHIRAJAN

*Indian Statistical Institute*

1. The object of this note is to prove that a bounded proper distribution can not be infinitely divisible (I.D.).

2. The following result of Polya (1949), will be used: A necessary and sufficient condition that a probability distribution should be bounded is that the definition of the characteristic function  $f(t)$  can be extended to complex values of the variable and this extension shows that  $f(t)$  is an entire function of exponential type. Moreover, if the distribution function is denoted by  $F(x)$  then the right and left extremities are given respectively by

$$\lim_{r \rightarrow +\infty} r^{-1} \log |f(-ir)| \quad \text{and} \quad -\lim_{r \rightarrow +\infty} r^{-1} \log |f(ir)|$$

**Theorem :** *A proper bounded distribution can not be I.D.*

*Proof :* If possible, let  $F(x)$  be a proper distribution bounded and I.D. In view of the fact that a random variable (r.v.)  $X$  is I.D. if and only if  $X-a$  is I.D., where 'a' is any real number, we may take the bounds as 0 and  $h$  ( $h > 0$ ). To show that such a distribution is not possible we will prove that  $h$  is necessarily infinite. As it is I.D., its characteristic function (c.f.)  $f(t)$  is represented by Gnedenko, B. and Kolmogorov, A. N. (1954).

$$\log f(t) = i\gamma t + \int \frac{e^{itu} - 1 - itu}{u^2} dK(u)$$

where  $\gamma$  is a constant and  $K(u)[K(-\infty) = 0]$  is a non decreasing function of bounded variation.

Further 
$$K(u) = \lim_{n \rightarrow \infty} K_n(u) \quad \text{for each } u$$

where 
$$K_n(u) = n \int_{-\infty}^u z^2 dF_n(z)$$

$F_n(z)$  being the distribution function (d.f.) corresponding to the c.f.  $[f(t)]^{1/n}$ . That this is a c.f. follows from the fact that  $F(x)$  is I.D. By Polya's Theorem  $F_n(x)$  is a bounded d.f. with lower bound zero.

$$\begin{aligned} \therefore K_n(u) &= 0 \quad \text{if } u \leq 0 \\ &= n \int_0^u z^2 dF_n(z) \quad \text{if } u > 0 \end{aligned}$$

$$\therefore K(u) = 0 \quad \text{for } u \leq 0$$

Here 
$$\log f(t) = i\gamma t + \int_0^\infty \frac{e^{itu} - 1 - itu}{u^2} dK(u)$$

The same representation goes over for  $t = ir$ ,  $r$  "real" as can be easily proved by means of the proof in Gnedenko, B. and Kolmogorov, A. N. (1954), and using simple properties of entire functions.

Hence by Polya's Theorem

$$\begin{aligned} h &= \overline{\lim}_{r \rightarrow +\infty} \frac{1}{r} \log |f(-ir)| \\ &= \overline{\lim}_{r \rightarrow +\infty} \left[ \gamma + \frac{1}{r} \int_0^\infty \frac{e^{ru} - 1 - ru}{u^2} dK(u) \right] \end{aligned}$$

Now,

$$\begin{aligned} &\frac{1}{r} \int_0^\infty \frac{e^{ru} - 1 - ru}{u^2} dK(u) \\ &> -\frac{r}{2} \int_0^\infty dK(u) \\ &> M.r \end{aligned}$$

Therefore,  $h = \infty$  unless  $K(u)$  is constant over  $(0, \infty)$  in which case the law is improper.

We are thankful to Dr. G. Kallianpur for suggestions and criticisms.

#### REFERENCES

- POLYA, G. (1949): Remarks on characteristic functions, *First Berkeley Symposium*, 115-123.  
 GNEDENKO, B. AND KOLMOGOROV, A. N. (1954): Limit distributions of sums of independent random variables (Trans. by K. L. Chung).

*Paper received : April, 1956.*



# A NOTE ON THE CONSTRUCTION OF ORTHOGONAL LATIN SQUARES

By NIKHILESH BHATTACHARYA

*Indian Statistical Institute, Calcutta*

Mann (1949) has given a method of constructing a set of  $\min_i (p_i^{e_i} - 1)$  mutually orthogonal  $s$ -sided latin squares when  $s$  is of the form  $s = \prod_{i=1}^k p_i^{e_i}$ , the  $p_i$ 's being different primes. Mann's method utilises properties of Galois fields. The following illustrates an alternative method of construction which is quite general.

2. Consider two latin squares, one 3-sided and the other 4-sided. Let the three-sided square be, say,

$$\begin{array}{ccc} 1 & 2 & 3 \\ 2 & 3 & 1 \\ 3 & 1 & 2 \end{array} \equiv L_1.$$

Associated with this are the following three latin squares  $L_2$ ,  $L_3$ , and  $L_4$ , obtained by adding 3, 6, and 9 respectively to each element of  $L_1$ :

$$L_2 \equiv \begin{array}{ccc} 4 & 5 & 6 \\ 5 & 6 & 4 \\ 6 & 4 & 5 \end{array}; \quad L_3 \equiv \begin{array}{ccc} 7 & 8 & 9 \\ 8 & 9 & 7 \\ 9 & 7 & 8 \end{array}; \quad L_4 \equiv \begin{array}{ccc} 10 & 11 & 12 \\ 11 & 12 & 10 \\ 12 & 10 & 11 \end{array}$$

Let the 4-sided latin square be, say,

$$\begin{array}{cccc} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \\ 3 & 4 & 1 & 2 \\ 4 & 1 & 2 & 3 \end{array}$$

3. If now we arrange the four 3-sided squares  $L_1$ ,  $L_2$ ,  $L_3$  and  $L_4$  in the way in which the numbers 1, 2, 3, and 4 respectively are occurring in the 4-sided square we arrive at the following 12-sided latin square:

$$\begin{array}{cccc} \begin{array}{ccc} 1 & 2 & 3 \\ 2 & 3 & 1 \\ 3 & 1 & 2 \end{array} & \begin{array}{ccc} 4 & 5 & 6 \\ 5 & 6 & 4 \\ 6 & 4 & 5 \end{array} & \begin{array}{ccc} 7 & 8 & 9 \\ 8 & 9 & 7 \\ 9 & 7 & 8 \end{array} & \begin{array}{ccc} 10 & 11 & 12 \\ 11 & 12 & 10 \\ 12 & 10 & 11 \end{array} \\ \begin{array}{ccc} 4 & 5 & 6 \\ 5 & 6 & 4 \\ 6 & 4 & 5 \end{array} & \begin{array}{ccc} 7 & 8 & 9 \\ 8 & 9 & 7 \\ 9 & 7 & 8 \end{array} & \begin{array}{ccc} 10 & 11 & 12 \\ 11 & 12 & 10 \\ 12 & 10 & 11 \end{array} & \begin{array}{ccc} 1 & 2 & 3 \\ 2 & 3 & 1 \\ 3 & 1 & 2 \end{array} \\ \begin{array}{ccc} 7 & 8 & 9 \\ 8 & 9 & 7 \\ 9 & 7 & 8 \end{array} & \begin{array}{ccc} 10 & 11 & 12 \\ 11 & 12 & 10 \\ 12 & 10 & 11 \end{array} & \begin{array}{ccc} 1 & 2 & 3 \\ 2 & 3 & 1 \\ 3 & 1 & 2 \end{array} & \begin{array}{ccc} 4 & 5 & 6 \\ 5 & 6 & 4 \\ 6 & 4 & 5 \end{array} \\ \begin{array}{ccc} 10 & 11 & 12 \\ 11 & 12 & 10 \\ 12 & 10 & 11 \end{array} & \begin{array}{ccc} 1 & 2 & 3 \\ 2 & 3 & 1 \\ 3 & 1 & 2 \end{array} & \begin{array}{ccc} 4 & 5 & 6 \\ 5 & 6 & 4 \\ 6 & 4 & 5 \end{array} & \begin{array}{ccc} 7 & 8 & 9 \\ 8 & 9 & 7 \\ 9 & 7 & 8 \end{array} \end{array}$$

4. If we start from another pair of latin squares, one ( $L'_1$ ) 3-sided and the other 4-sided, which are orthogonal to those used in above, we shall get a 12-sided latin square orthogonal to that obtained above. For when the two latin squares are superposed, any pair of numbers say 4 in the first square and 8 in the second will coincide only when the subsquare  $L_2$  (containing 4) of the first latin square falls upon subsquare  $L'_3$  (containing 8) of the second, which they do only once, in virtue of the orthogonality of the two 4-sided latin squares used. Also when  $L_2$  and  $L'_3$  are superimposed, 4 in  $L_2$  and 8 in  $L'_3$  come together in the one cell whose position is similar to the one cell (in virtue again of orthogonality) where 1(=4-3) in  $L_1$  and 2(=8-2×3) in  $L'_1$  coincide when  $L_1$  and  $L'_1$  are superposed.

5. If  $s = p_1^{e_1} p_2^{e_2}$ , we get by the methods of Bose and Stevens outlined by Mann (1949), a set of  $(p_1^{e_1} - 1)$  orthogonal squares of side  $p_1^{e_1}$  and another set of  $(p_2^{e_2} - 1)$  orthogonal squares of side  $p_2^{e_2}$ . We can go on taking from these two sets one square of side  $p_1^{e_1}$  and another of side  $p_2^{e_2}$  and "combining" them in the manner indicated to get  $\min(p_1^{e_1}, p_2^{e_2}) - 1$  orthogonal  $s$ -sided latin squares. This can be easily extended to the case  $s = \prod_{i=1}^k p_i^{e_i}$ .

## REFERENCE

MANN, H. B. (1949): *Analysis and Design of Experiments : Analysis of variance and Analysis of variance designs*. Dover Publications, New York.

*Paper received : October, 1955.*

# A NEW DISCRETE DISTRIBUTION

By AYODHYA PRASAD

*Agricultural Research Institute, Sabour, India*

The following is an interesting discrete distribution whose spectrum consists of values 1, 2, ...,  $\infty$ .

$$\xi = 1, 2, 3, \dots$$

$$\zeta_{\xi} = \frac{2\lambda(\lambda+1)}{(\lambda+\xi-1)(\lambda+\xi)(\lambda+\xi+1)}, \lambda > 0,$$

$$\sum_{\xi=1}^{\infty} \zeta_{\xi} = \lambda(\lambda+1) \sum_{\xi=1}^{\infty} \frac{2}{(\lambda+\xi-1)(\lambda+\xi)(\lambda+\xi+1)} = \frac{\lambda(\lambda+1)}{\lambda(\lambda+1)} \equiv 1$$

$\zeta_{\xi}$  is the probability that the random variable  $\xi$  assumes integral values in the region  $(1, \infty)$  and  $\lambda$  is a parameter.

The first moment about the origin which is also the population mean  $m$ , is

$$\mu'_1 = 1 + \lambda.$$

The second moment about the mean is infinite and the determination of the standard deviation is out of question. However, the mean deviation of the variable can be shown to be given by either

$$\delta_m = \frac{2\lambda(\lambda+1)}{2\lambda+1}, \lambda = n,$$

where  $n$  is an integer or by

$$\delta_m = \frac{4\lambda(\lambda+1)(n+1)}{(\lambda+n+1)(\lambda+n+2)}, \lambda = n + \theta,$$

where  $n$  is an integer or zero and  $0 < \theta < 1$ .

For smaller values of  $\lambda$ , there is a rapid or steep fall in the values of  $\zeta_{\xi}$  with higher values of  $\xi$  and for larger values of  $\lambda$ , there is a gradual or slow fall in values of  $\zeta_{\xi}$  with higher values of  $\xi$ . It is worthwhile contrasting these probabilities with those of a Poisson distribution in which there is only a fall in values of the probability with higher values of the variable when the parameter is smaller but there is rise and fall both in them for larger values of the parameter.



If  $l$ , the length of the needle and  $d$ , the distance between two consecutive lines in the Buffon's Needle Problem (Uspensky, 1937) be made functions of  $\lambda$  and  $\xi$  as shown below

$$l = \lambda(\lambda+1)$$

$$\pi d = (\lambda+\xi-1)(\lambda+\xi)(\lambda+\xi+1),$$

then  $\zeta_\xi$  under suitable condition gives the probability of the needle intersecting one of the lines when thrown at random on the board.

Since  $l < d$  in the said problem, it can be easily shown that the suitable condition is satisfied if  $\lambda \geq 2$ . Then if the Buffon's Needle Problem is so designed that  $l$  and  $d$  cease to vary independently, obviously then  $\zeta_\xi \left( = \frac{2l}{\pi d} \right)$  which is nothing but the probability of the needle intersecting one of the lines, gives a probability distribution with respect to  $\xi$  when  $l$  and  $d$  are replaced by  $\lambda$  and  $\xi$  as above.

#### REFERENCE

USPENSKY, J. V. (1937) : *Introduction to Mathematical Probability*, McGraw Hill and Sons, New York.

*Paper received: December, 1955.*

# AN EXPERIMENTAL METHOD FOR OBTAINING RANDOM DIGITS AND PERMUTATIONS

By JOHN E. WALSH

*Lockheed Aircraft Corporation, California, U.S.A.*

This paper presents an easily applied method for obtaining small numbers of random binary digits and random permutations. The procedure consists in flipping ordinary minted coins and combining the results of the flips in an appropriate manner. Digits and permutations obtained according to the method of this paper can be considered sufficiently random for any practical application. It appears likely that these digits and permutations are much more nearly random than most of those now available in printed tables. Moreover, any possibility of bias from misuse of tables is avoided. The method presented is particularly suitable for use with respect to experimental designs. Only a few random permutations are ordinarily required for a given experimental design.

## INTRODUCTION AND DISCUSSION

The outputs of probability processes which produce random digits and random permutations have many uses. Two fields of application are in the design of experiments and the theory of games. For many cases, the number of digits or permutations needed for an application is small. Thus easily applied experimental procedures for producing small numbers of digits and permutations which are sufficiently random for any practical application may be of interest. This paper presents such a procedure. The present section is devoted to a discussion of the basic concepts involved and to consideration of the advantages and disadvantages of the method advocated. An explicit specification of the procedure is presented in the next section.

First let us consider what is meant by the statement that a set of digits is random. This statement implies that the values of the digits are observations produced by a process which satisfies certain probability conditions. Once the values of the digits are recorded, they are merely a set of numbers. The probability properties of the process which produced these numbers are the items of interest. One of these probability properties is that the observations produced by the process are statistically independent. The other property is that the probability of an observation having any specified value equals the reciprocal of the number base for the type of digits being considered. As an example, if decimal digits are used (base 10), the probability that the value of an observation equals any specified one of the digits 0, 1, ..., 9 has the value  $1/10$ . As another example, consider binary digits (base 2); then the probability that an observation has any specified one of the values 0, 1 equals  $1/2$ .

Next consider what is meant by a random permutation. In this paper the quantities permuted are the numbers 1, 2, ...,  $r$ . The problem is to select a permutation of these numbers in such a manner that a certain probability condition is satisfied. This probability condition requires that the selection be made so that all possible permutations have the same probability of being chosen. For example, let  $r = 3$ . Then the six possible permutations are 123, 132, 213, 231, 312, 321 and each has the probability  $1/6$  of being selected. The usual method of obtaining random permutations consists in using random digits for the selection of permutations.

Let us assume that processes for producing truly random digits and permutations are available. One way to make use of these processes is only to produce digits and permutations when they are needed. A more convenient way, however, is to produce large numbers of digits and permutations at one time, record the results, and then use these results as they are needed. If the second way is used, two requirements should be satisfied. First, the order in which the recorded digits and permutations are used should be specified without any knowledge of the recorded values. It would seem preferable to use them in the order in which they were produced by the processes. Second, the recorded digits and permutations should only be used once. If these two rules are not followed, the digits and permutations are not necessarily the outputs of truly random processes.

At present, the random digits ordinarily available are in the form of printed tables. That is, they are considered to represent the recorded results of truly random processes. Use of these tables for obtaining small numbers of random digits is open to several sources of error. First, one or both of the two rules stated in the preceding paragraph are often violated by the method used in selecting the digits from the table. Second, the process used to produce the digits may not have been sufficiently random. Finally, use of the same table by two or more persons may destroy the independence of their results.

The safe way to use a random digit table is to specify the order in which the digits will be utilized before any knowledge of the actual values is available and discard the table when all the digits have been used. Other procedures may lead to noticeable bias. In particular, the set of digits used should not represent a personal choice of the experimenter. Otherwise the experiment is open to criticism. Namely, the results depend on the particular set of digits selected and there may be no good reason for choosing this set rather than some other set which would have yielded entirely different results.

The two principal methods which have been used to obtain random digits are expansion of transcendental numbers and construction of physical processes. The motivation for the expansion method is the ergodic theorem. The accuracy of the physical process method is based on technical considerations. However, examination has shown that the Ergodic theorem may not be even approximately satisfied for the first part of an expansion (say, the first  $10^{20}$  digits). No matter how much care is exercised in the construction of a physical process, biases can enter and disappear while the process is in operation. Thus it appears likely that most of the random digit tables now available only represent the results of a random process to a first approximation. The randomness of existing tables was checked by statistical tests applied to the values of the digits. Here the reasoning is that the digits are sufficiently random if there is no strong reason to believe otherwise. This approach is of a negative nature and does not prove that a given set of digits is very nearly random. It merely indicates that this might be the case.

The procedure presented in this paper starts with the outputs of a physical process. This process consists in flipping ordinary minted coins. These outputs are then manipulated to yield a set of digits. The digits produced by this method can be rigorously proved to be sufficiently random if two obviously acceptable assumptions are satisfied. These assumptions are

- (i) For each flip of a coin, the probability of obtaining a "head" lies somewhere between  $1/4$  and  $3/4$ .
- (ii) The flips of the coins are statistically independent.



## A METHOD FOR OBTAINING RANDOM DIGITS AND PERMUTATIONS

If any reasonable amount of care is exercised in choosing the coins to be flipped and in flipping them, assumptions (i) and (ii) should be acceptable with virtual certainty.

The random digits produced by the procedure of the paper are binary rather than decimal. That is, each digit has one of the values 0 or 1. A set of random binary digits can be considered to represent the recorded results of the independent flips of an "ideal" coin which has one side denoted by 1 and the other side by 0. An "ideal" coin has the property that the probability of obtaining a specified side equals  $1/2$ . With this interpretation, there should be little difficulty in understanding the use of random binary digits.

For some situations, random permutations are much more useful than random digits. This is the case, for instance, if the problem is to randomly assign a set of objects to a set of positions. This type of problem arises frequently in the design of experiments. Permutations of 16 numbers ( $r = 16$ ) have been found to be useful for performing the randomizations employed in experimental designs (Cochran and Cox 1950). Random binary digits are easily used to obtain random permutations of 16 numbers. The next section contains explicit instructions for obtaining random permutations of the numbers 1, 2, ..., 16 from the binary digits produced by the method of the paper. For obtaining random permutations of 16 numbers, random binary digits are much more suitable than random decimal digits.

In deriving the method stated in this paper, the theoretical results presented in (Walsh, 1949) were used. The stated procedure yields a set containing at most 64 binary digits at each application. Let assumptions (i) and (ii) be satisfied and consider the accuracy of the method advocated. Use of the theory shows that the probability of any relation based on a set of 64 or less binary digits obtained by the procedure of the paper never differs from the hypothetical value by more than 0.8% of that value. Moreover, this deviation from the corresponding probability value for a truly random set of binary digits is as large as 0.8% only in the most extreme cases. Thus random binary digits obtained by the method of this paper should be suitable for any practical application.

The principal disadvantage of the method presented would appear to be the amount of work required in obtaining the binary digits. To obtain 64 digits, 650 coin flips are needed. More than this number of flips are ordinarily required to produce a permutation of 16 numbers. If ten coins are available, a suggested procedure would be to flip these ten coins, record the result (see the next section for a statement of the method), flip the ten coins again, record the result, etc. Under favourable conditions, production of 64 digits should not take more than half an hour. Moreover, the procedure of recording digits for future use allows spare time and slack periods to be utilized in producing random binary digits.

Some valuable properties of the method of the paper have already been mentioned. Let us consider a summary of the principal advantages:

1. The digits and permutations produced are proved to be sufficiently random for any practical application if two intuitively acceptable assumptions hold.
2. Eliminates the necessity of having random digit and/or random permutation tables.

The procedure used to obtain the digits and permutations automatically yields the additional advantages:

3. Avoids biases which can arise from misuse of tables.
4. Avoids criticisms which can be applied to subjective use of tables.

5. Random digits and permutations produced are independent of those obtained by anyone else.

Whether these advantages outweigh the disadvantage of extra work depends on the particular situation being considered.

#### STATEMENT OF METHOD

The method is based on flips of ordinary minted coins. It is not necessary to use the same coin for each flip nor is it necessary to use a different coin for each flip. In fact, the selection of coins for the flips is arbitrary except that no coin should be so malformed that there is any doubt about the validity of assumption (i). In flipping the coins, each coin should be flipped separately. The coin should be tossed high into the air, preferably revolving, and should land on a hard surface. If these rules for selection of coins and flipping are followed, assumptions (i) and (ii) should be acceptable beyond any doubt.

The method is designed to produce at most 64 binary digits at a time. Let  $n$  represent the actual number of digits to be produced ( $n \leq 64$ ). The procedure consists in obtaining  $n+1$  sets of ten flips. The first ten flips represent the first set, the second ten flips the second set, etc. For the first set of ten flips, consider whether the number of "heads" obtained was even or odd. Next consider the remaining  $n$  sets of ten flips. If the number of "heads" obtained for the first set was even, for each of the remaining sets an even number of "heads" will be denoted by 0 and an odd number of "heads" by 1. If the number of "heads" for the first set was odd, in each of the other sets an even number of "heads" will denote 1 and an odd number 0. For the second, third, ...,  $(n+1)$ st sets, record either a 0 or 1 according to this rule; the order for the recordings is the same as for the sets. The resulting 0's and 1's represent the desired set of  $n$  binary digits arranged in the order in which they are to be used. If more than 64 digits are required, the procedure stated in this paragraph is repeated.

*Example of Application.* As a numerical illustration of the method presented in the preceding paragraph, let  $n = 5$  and suppose that the six sets of ten flips (arranged in order) yielded the following results:

set	1	2	3	4	5	6
no. of "heads"	4	6	5	8	6	3

Since the number of "heads" in the first set is even, for all the other sets an even number of "heads" furnishes a 0 and an odd number a 1. Thus the resulting five binary digits (arranged in order) are

0, 1, 0, 0, 1.

Now, let us discuss the procedure for obtaining random permutations from random binary digits. Only random permutations of 16 numbers will be considered.

The problem is that of assigning the numbers 1, 2, ..., 16 to an ordered set of sixteen positions. This is to be done so that all sixteen numbers have the same probability of being assigned to the position. After the first position has been filled, the remaining fifteen numbers have the same probability of being assigned to the second position; after the first and second positions have been filled, the remaining fourteen numbers have the same probability of being assigned to third position; etc., until all sixteen positions have been filled.



## A METHOD FOR OBTAINING RANDOM DIGITS AND PERMUTATIONS

Arrangement of the sixteen numbers according to the order of the positions they occupy furnishes a random permutation. For example, the permutation obtained might be

5, 9, 8, 16, 2, 4, 10, 1, 3, 6, 15, 11, 7, 13, 12, 14.

Random binary digits are used to assign the numbers to the positions in such a manner that each eligible number has the same probability of being assigned to a specified position.

Consider a set of random binary digits arranged in the order in which they are to be used. There are sixteen possible combinations for the values of the first four digits. Let the following correspondence exist between these sixteen combinations and the numbers 1, 2, ..., 16:

1	↔	0000,	9	↔	1000,
2	↔	0001,	10	↔	1001,
3	↔	0010,	11	↔	1010,
4	↔	0011,	12	↔	1011,
5	↔	0100,	13	↔	1100,
6	↔	0101,	14	↔	1101,
7	↔	0110,	15	↔	1110,
8	↔	0111,	16	↔	1111.

Then the first set of four binary digits furnishes the number to be placed in the first position of the permutation. Next consider the second set of four binary digits. This yields another number. If this number is different from that in the first position, it is used for the second position. If the number is the same as for the first position it will not be used in this permutation. However, this number is not discarded; it is used to fill the first position of a second permutation.

Now consider the third set of four digits. If the number corresponding to this set has not yet appeared in the first permutation, use this number to fill the first vacant position. For instance, if the first two positions are occupied, it would be placed in the third position; if only the first position is filled this number would be placed in the second position. When the number has already appeared in the first permutation, it is considered for use in the second permutation. If no numbers have been used in the second permutation, it will occupy the first position. When the first position has been filled, it will be used in the second position if different from the number in the first position. When this number is the same as that in the first position of the second permutation, it is used to fill the first position of a third permutation. In general, the number corresponding to the set of four binary digits being used is first considered for use in the first permutation. If it is different from the numbers which have already appeared, it is placed in the first vacant position. If this number is duplicated in the first permutation, it is considered for use in the second permutation. This procedure is continued until the number can be used in the first vacant position of some permutation. This might be the first position of a new permutation.

Use of the method just outlined guarantees that every number yielded by a set of four binary digits will appear in some permutation. Thus, in the long run, nothing is wasted. As additional digits are utilized, the positions in the permutations are filled up. When all sixteen of its positions are filled, a permutation is recorded for future application. The order in which the permutations are used is the order in which they are recorded.



It can be verified that each permutation obtained according to the above procedure is a random permutation. However, the permutations are not independent. This is the price paid for using a method which utilizes all the numbers furnished by the random digits. Fortunately, in many cases this lack of independence causes little difficulty. These are situations where only a few random permutations are required (e.g. in experimental designs). Then several separate constructions of random permutations are employed. For an application, each permutation used is taken from a different construction. The permutations applied are then both random and independent.

*Numerical Example.* Let us consider an explicit illustration of the procedure for obtaining random permutations of 16 numbers by use of random binary digits. Suppose that the first 64 binary digits (in order) are

1001 0111 0011 0100 0010 1100 1110 1011  
0010 1101 0010 1111 0100 1010 0101 1100

where the first digit of the second row follows the last digit of the first row in the ordering. On the basis of these 64 digits, the positions of the permutations are occupied as follows:

1st Permutation: 10, 8, 4, 5, 3, 13, 15, 12, 14, 16, 11, 6  
2nd Permutation: 3, 5, 13  
3rd Permutation: 3.

The suitability of the method of obtaining random binary digits follows directly from the results of (Walsh, 1949) combined with assumptions (i) and (ii). To avoid the introduction of concepts which would require much explanation, derivations will not be presented in this paper. The principal result has already been stated in the preceding section. Namely, the probability of any relation based on a set of 64 or fewer digits produced by the procedure of the paper never differs from the hypothetical value by more than 0.8%. It is to be observed that a random permutation of 16 numbers is based on exactly 64 random binary digits. The derivations are very straightforward; a mathematical statistician should have little difficulty in verifying the principal result if he applies the theory of (Walsh, 1949) and assumptions (i) and (ii) to the method of the paper.

Although the procedure for experimentally producing binary digits has been limited to the flipping of coins, the same type of digit manipulations can be applied to various other physical processes. Suitable selection of the physical process combined with a mechanization of the digit manipulation procedure could yield a method for producing random digits which is both rapid and easily applied.

#### REFERENCES

- COCHRAN, W. G. and COX, G. M. (1950): *Experimental Designs*, John Wiley and Sons, 414-21.  
WALSH, JOHN E. (1949): Concerning compound randomization in the binary system. *Ann. Math Stat.*, 20, 580-89.

*Paper received : July, 1955.*

# ON ESTIMATING PARAMETRIC FUNCTIONS IN STRATIFIED SAMPLING DESIGNS

By DES RAJ

*Indian Statistical Institute, Calcutta*

## 1. THE PROBLEM CONSIDERED

In the usual theory of stratified sampling designs, we are generally interested in estimating the mean value or aggregate of a character for a finite population. We may, however, come across situations in which the object is to estimate certain linear functions of the stratum means. As a case in point, we may require to estimate the area under food crops not only for a province as a whole but also for a particular group of strata within the province where for instance famine might have occurred or it is feared that much area is being diverted to crops of one type at the expense of crops of a more beneficial type. As another example, we may wish to estimate not only the average consumption of rice by all the inhabitants in a city but also the average consumptions by the primarily rice-eating classes and primarily wheat-eating classes so that Government may be in a position to find out how much rice should be procured for the city and in what manner it should be rationed to the different classes.

## 2. ESTIMATION OF PARAMETRIC FUNCTIONS

In general, if the population consists of  $k$  strata of sizes  $N_j (j = 1, 2, \dots, k)$  and mean values

$$\bar{Y}_j (j = 1, 2, \dots, k), \quad \dots \quad (2.1)$$

we are interested in estimating  $r \leq k$  linear functions

$$L_i = \sum_{j=1}^k l_{ij} \bar{Y}_j \quad (i = 1, 2, \dots, r) \quad \dots \quad (2.2)$$

of the stratum means, the matrix of co-efficients  $(l_{ij})$  being known.

For a particular distribution  $n_j (j = 1, 2, \dots, k)$  of the total sample of size  $n$ , obviously the best unbiased estimate of  $L_i$  is

$$\hat{L}_i = \sum_{j=1}^k l_{ij} \bar{y}_j \quad (i = 1, 2, \dots, r) \quad \dots \quad (2.3)$$

where  $\bar{y}_j$  is the sample mean based on  $n_j$  observations in the  $j$ -th stratum.

These estimates are best (Basu, 1952) in the sense that for any convex (downwards) loss function they are admissible i.e., no other estimators having a uniformly smaller risk function exist. Also we have

$$V(\hat{L}_i) = \sum l_{ij}^2 \sigma_j^2 \left( \frac{1}{n_j} - \frac{1}{N_j} \right)$$

$$\sigma_j^2 = \frac{\sum_{k=1}^{N_j} (y_{jk} - \bar{Y}_j)^2}{N_j - 1} \quad \dots \quad (2.4)$$

where

is the variance of the  $j$ -th stratum. Further the best (in the sense stated above) unbiased quadratic estimates of  $V(\hat{L}_i)$  are given by

$$\hat{V}(\hat{L}_i) = \sum l_{ij}^2 s_j^2 \left( \frac{1}{n_j} - \frac{1}{N_j} \right)$$

where

$$s_j^2 = \frac{\sum_{k=1}^{n_j} (y_{jk} - \bar{y}_j)^2}{n_j - 1} \quad \dots (2.5)$$

### 3. OPTIMUM ALLOCATION OF SAMPLING UNITS

An important question emerges: How should the total sample be distributed among the different strata? Obviously, there would not be a single answer to this question. In fact, the answer will depend on what the sample is desired to achieve. In what follows we shall consider some approaches to the problem.

3.1. *Minimisation of cost plus loss*: If the results obtained from the sample are going to form the basis of some practical action, we may be able to calculate in monetary terms the 'loss' that will be incurred in a decision through an error of amount  $d$  in the estimate. For example, if this loss be  $\mu_i d^2$  (cf. Yates, p. 292) and the estimate be unbiased, the average loss in a series of samples of the same type and size will be  $\mu_i V(\hat{L}_i)$ . The purpose in taking the sample may be to diminish the sum of the total expected loss

$$L = \sum_{i=1}^r \mu_i V(\hat{L}_i) \quad \dots (3.1.1)$$

and the total cost (cf. Kitagawa, p. 338)

$$C = \sum c_j n_j^g \quad (g > 0). \quad \dots (3.1.2)$$

Or, for a fixed cost given by (3.1.2), the object may be to diminish the expected loss given by (3.1.1). The cost function used here is more general than the usual cost function

$$C = \sum c_j n_j.$$

In the former case, the function to be minimised is

$$G = \sum c_j n_j^g + \sum q_j \sigma_j^2 \left( \frac{1}{n_j} - \frac{1}{N_j} \right) \quad \dots (3.1.3)$$

where

$$q_j = \sum_i \mu_i l_{ij}^2. \quad \dots (3.1.4)$$

The stationary value of  $G$  is given by

$$n_j^{g+1} = \frac{q_j \sigma_j^2}{g c_j} \quad \dots (3.1.5)$$



## PARAMETRIC FUNCTIONS IN STRATIFIED SAMPLING DESIGNS

It is easy to verify that the stationary value is an actual minimum of the function. In the latter case, the solution is given by

$$n_j = \mu \left( \frac{q_j \sigma_j^2}{g c_j} \right)^{\frac{1}{\theta+1}} \quad \dots \quad (3.1.6)$$

where

$$\mu = \frac{C^{1/\theta}}{\left[ \sum (q_j \sigma_j^2 / g)^{\frac{\theta}{\theta+1}} / c_j^{\theta} \right]^{1/\theta}} \quad \dots \quad (3.1.7)$$

**3.2. Minimisation of cost:** As an alternative approach to the problem, we may consider the survey to be useful if the parametric functions  $L_i$  ( $i = 1, 2, \dots, r$ ) are estimated with some desired variances  $a_i$  ( $i = 1, 2, \dots, r$ ). Generally approximate value of  $L_i$ 's are known on the basis of some previous survey and  $a_i$ 's are determined by the requirement that the standard errors of the estimates are some specified percentages of the mean values known approximately. In such a case, we have to allocate the total sample to the different strata in such a way that the cost of the survey is made smallest. We have then to minimise

$$f(n_1, n_2, \dots, n_k) = \sum c_j n_j^g$$

subject to the conditions  $\phi_i = V(\hat{L}_i) = a_i$  ( $i = 1, 2, \dots, r$ ).

The stationary values are given by the equations

$$m_j^{g+1} = g c_j [\sigma_j^2 \sum_i \lambda_i l_{ij}^2]^{-1} \quad (j = 1, 2, \dots, k) \quad \dots \quad (3.2.1)$$

$$\sum_j l_{ij}^2 m_j \sigma_j^2 = a_i + \sum_j l_{ij}^2 \sigma_j^2 / N_j \quad (i = 1, 2, \dots, r) \quad \dots \quad (3.2.2)$$

where  $\lambda$ 's are Lagrange's undetermined multipliers and  $m_j$  is the reciprocal of  $n_j$ .

These equations are not algebraic in the variables involved. It is, therefore, not possible to provide any explicit mathematical solutions for the general case. One has to solve these equations by iterative processes. For example, if  $r = 2$ , an approximate solution can easily be obtained by Newton-Raphson method. In this method, we choose  $\lambda_1$  so that the difference between the values of  $\lambda_2$  calculated from the two equations obtained from (3.2.2) by substituting for  $m_j$  from (3.2.1) is small and positive. We find another value of  $\lambda_1$  for which this difference is small and negative. By simple interpolation values of  $\lambda_1$  and  $\lambda_2$  are obtained which lead to the optimum allocation.

One may like to verify that the stationary point obtained is an actual minimum of  $f(n_1, n_2, \dots, n_k)$ . Putting  $f + \lambda_1 \phi_1 + \lambda_2 \phi_2 + \dots + \lambda_r \phi_r = F$  and denoting by  $h^{ij}$  the second partial derivative of  $h$  w.r.t.  $n_i$  and  $n_j$ , the condition is that the restricted Hessian (Chaundy, 1935)

$$\begin{vmatrix} F^{11} & F^{12} & \dots & F^{1k} & \phi_1^1 & \phi_1^2 & \dots & \phi_1^r \\ F^{21} & F^{22} & \dots & F^{2k} & \phi_2^1 & \phi_2^2 & \dots & \phi_2^r \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ F^{k1} & F^{k2} & \dots & F^{kk} & \phi_k^1 & \phi_k^2 & \dots & \phi_k^r \\ \phi_1^1 & \phi_1^2 & \dots & \phi_1^k & 0 & 0 & \dots & 0 \\ \phi_2^1 & \phi_2^2 & \dots & \phi_2^k & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \phi_r^1 & \phi_r^2 & \dots & \phi_r^k & 0 & 0 & \dots & 0 \end{vmatrix} \quad \dots \quad (3.2.3)$$

and its principal minors should have the sign  $(-1)^r$ . It may also be noted that the solution so obtained is formally equivalent to the following :

If  $(n'_1, n'_2, \dots, n'_k)$  is a minimum of  $f$  subject to the  $r$  conditions

$$\phi_i = V(\hat{L}_i) = a_i \quad (i = 1, 2, \dots, r)$$

it is also a minimum or maximum of  $\phi_s$  subject to the  $r$  conditions

$$f = f(n'_1, n'_2, \dots, n'_k),$$

$$\phi_t = a_t \quad (t = 1, 2, \dots, s-1, s+1, \dots, r)$$

according as  $\lambda_s$  defined by the  $k$  equations

$$\frac{\partial f}{\partial n_p} + \sum_{i=1}^{s-1} \lambda_i \frac{\partial \phi_i}{\partial n_p} + \lambda_s \frac{\partial \phi_s}{\partial n_p} + \sum_{t=s+1}^r \lambda_t \frac{\partial \phi_t}{\partial n_p} = 0 \quad (p = 1, 2, \dots, k)$$

is positive or negative. Moreover, the minimum or maximum value of  $\phi_s$  is  $a_s$ . This means that for a certain cost and some specified variances of any  $r-1$  estimates, we are minimising the variance of the  $r$ -th estimate.

**3.3. Minimisation of variances :** Another type of requirement may be that relative precisions of the different estimates be in some assigned ratios. As a particular case, it may be desired that the coefficients of variation of the different estimates be all equal, the common value of the coefficients of variation necessarily depending on the cost of the survey. In such a situation the variance of one of the variates will be minimised for fixed cost and for stipulated relations between the variances. As an example of such a situation these estimates may be required by different agencies (like State Governments), and if the precision of an estimate is judged by its variance, we may distribute the total sample size so that the relative precisions of the different estimates are all equal.

If the relative precisions of the estimates are governed by

$$V(\hat{L}_1) = v_2 V(\hat{L}_2) = \dots = v_r V(\hat{L}_r) \quad \dots (3.3.1)$$

and the cost is given by

$$\sum c_j n_j^q = C \quad \dots (3.3.2)$$

the optimum allocation is the solution of the equations

$$n_j^{q+1} = \frac{\sigma_j^2}{\lambda_{r+1} g c_j} \left[ l_{ij}^2 \sum_{i=1}^r \lambda_i - \sum_{i=1}^r \lambda_i v_i l_{ij}^2 \right], \quad (j = 1, 2, k; \lambda_1 = 1) \quad \dots (3.3.3)$$

along with (3.3.2) and the  $r-1$  equations given by (3.3.1).

Like the equations considered in the previous section, it is not possible to offer explicit mathematical solutions of these equations. They have to be solved by approximate methods. In particular, if  $r = 2$ , an approximate solution can be easily obtained using Newton-Raphson method on the lines indicated before.

# PARAMETRIC FUNCTIONS IN STRATIFIED SAMPLING DESIGNS

## 4. STRATIFICATION AFTER SELECTION

Sometimes it happens that the frame for the entire population is available but frames for individual strata are not known. In such a situation, since we cannot sample from individual strata, a simple random sample of size  $n$  is taken from the entire population. When the sample data have been collected the units are assigned to the strata by means of the information obtained about them. The best unbiased estimates are given by (2.3) as before and the expected variances are

$$E[V(\hat{L}_i)] = E \sum_{j=1}^k l_{ij}^2 \frac{\sigma_j^2}{n_j} = \sum_{j=1}^k l_{ij}^2 \frac{\sigma_j^2}{N_j} \quad \dots (4.1)$$

Using Stephan's (1945) result

$$E\left(\frac{1}{n_j}\right) = \frac{1}{nw_j} - \frac{1}{n^2 w_j} + \frac{1}{n^2 w_j^2} \quad \dots (4.2)$$

approximately where  $w_j$  is the relative size of the  $j$ -th stratum, we have

$$E[V(\hat{L}_i)] = \sum_j l_{ij}^2 \sigma_j^2 \left\{ \frac{1}{nw_j} - \frac{1}{n^2 w_j} + \frac{1}{n^2 w_j^2} \right\} - \sum_j l_{ij}^2 \frac{\sigma_j^2}{N_j} \quad \dots (4.3)$$

## 5. AN EXAMPLE

We now give an example to illustrate the various methods stated above. The object is to estimate the average area under wheat per village and the difference in the averages of the two strata, comprising villages with agricultural area below and above 1500 bighas respectively, for Ghaziabad *tahsil* in Uttar Pradesh (India). Data obtained from a previous census (Sukhatme, 1954) are given below in Table 1. We will assume that the stratum sizes  $N_i$  are exact as obtained from the previous census. The rest of the material will be used as supplementary information for improving the design of the current survey.

TABLE 1. RESULTS OF A PRELIMINARY CENSUS

strata	agr. area in bighas	no. of villages $N_i$	average area under wheat	standard deviation of area under wheat
(1)	(2)	(3)	(4)	(5)
1	0- 500	63	112	56
2	501-1500	199	277	116
3	1501-2500	53	558	186
4	2501 and above	25	960	361

Let the cost function be simply the number of villages and the expected loss be the variance of the estimate. The linear functions to be estimated are

$$L_1 = \frac{1}{N} \sum N_j \bar{Y}_j,$$

$$L_2 = \frac{N_1}{N_1+N_2} \bar{Y}_1 + \frac{N_2}{N_1+N_2} \bar{Y}_2 - \frac{N_3}{N_3+N_4} \bar{Y}_3 - \frac{N_4}{N_3+N_4} \bar{Y}_4.$$



(a) For a fixed cost of 34 villages, the optimum allocation minimising the total loss is given in column (2) of Table 2.

TABLE 2. OPTIMUM ALLOCATIONS IN DIFFERENT SITUATIONS

strata	case (a)	Neyman's allocation	case (b)	case (c)
(1)	(2)	(3)	(4)	(5)
1	2	3	3	2
2	7	17	19	10
3	8	7	20	11
4	17	7	18	11

The total loss in this case is 3903. The coefficients of variation of the estimates  $\hat{L}_1$  and  $\hat{L}_2$  are 8.27 percent and 12.46 percent respectively. If we had used Neyman's allocation based on the estimation of the mean only, the total loss would be 4292 and the individual coefficients of variation would be 6.29 percent and 13.77 percent respectively.

(b) If it be desired that the c.v. for  $\hat{L}_1$  be 5 percent and that for  $\hat{L}_2$  be 7.5 percent, we have to minimise the cost subject to the conditions

$$V(\hat{L}_1) = 289.77,$$

$$V(\hat{L}_2) = 1136.60.$$

The optimum allocation is given in column (4) of Table 2.

(c) In case it is desired that the c.v. of  $\hat{L}_1$  be approximately 2/3 c.v. ( $\hat{L}_2$ ) and that the cost of the survey be 34 villages we minimise  $V(\hat{L}_1)$  subject to the conditions

$$\sum n_j = 34,$$

$$V(\hat{L}_1) = .25 V(\hat{L}_2).$$

The allocation is given in column (5) of Table 2. It is found that

$$\text{c.v. } (\hat{L}_1) = 7.11 \text{ percent,}$$

$$\text{c.v. } (\hat{L}_2) = 10.77 \text{ percent.}$$

#### REFERENCES

- BASU, D. (1952): On symmetric estimators in point estimation with convex weight functions. *Sankhyā* 12, 45-52.
- CHAUNDY, T. (1935): *The Differential Calculus*, Oxford University Press, Oxford.
- COCHRAN, W. G. (1953): *Sampling Techniques*, John Wiley and Sons, New York.
- KITAGAWA, T. (1955): Some contributions to the design of sample surveys. *Sankhyā*, 14, 317-362.
- STEPHAN, F. F. (1945): The expected value and variance of the reciprocal and other negative powers of a positive Bernoullian variate, *Ann. Math. Stat.* 16, 50-61.
- SUKHATME, P. V. (1954): Sampling theory of surveys with applications. *Ind. Soc. Agr. Stat.*, New Delhi, 121.
- YATES, F. (1949): *Sampling Methods for Censuses and Surveys*, Charles Griffin and Company Ltd., London.

Paper received : May, 1954.

# A NOTE ON ESTIMATION OF VARIANCE COMPONENTS IN MULTISTAGE SAMPLING WITH VARYING PROBABILITIES

By J. ROY

*Indian Statistical Institute, Calcutta*

## 1. INTRODUCTION AND SUMMARY

In large scale surveys multistage sampling procedure using different probabilities of selection for different units at any particular stage have been used very often. In the Indian National Sample Survey (NSS), for instance, such a scheme is used for selection of units within a stratum. If in the first stage more than one unit is chosen with replacement, the standard error (of the estimate of the mean or of the total) can be easily computed from the variation between the different estimates that can be computed, one based on each of the chosen first stage unit. However, in order to be able to develop a suitable sampling scheme, we require not merely an estimate of the overall sampling error, but also a knowledge of how the error depends on the adjustable parameters at the disposal of the sampler. Cochran (1939) has shown how in the case of multistage simple sampling this leads to a problem of analysis of the total variation into different stage components. The corresponding problem when units are chosen with different probabilities (but with replacement) at each stage is dealt with in this paper. The total variation is split up into different meaningful components depending on the type of sampling used, and unbiased estimators for these components are derived.

## 2. THE SAMPLING SCHEMES

Here we shall consider a three-stage sampling scheme, but the method used is quite general and can be directly extended for any number of stages. We shall further assume that in every stage sampling is with replacement and that in the third stage units are chosen with equal probabilities. This scheme of sampling (with slight modifications) was used for selection of the ultimate unit (household) within a stratum in the first few rounds of the Indian National Sample Survey where within a stratum a tehsil served as a first-stage unit, villages within the tehsil as second stage units and households within a village as the ultimate third stage unit. Tehsils and villages were chosen with different probabilities, generally proportional to population or area and households within a village were selected with equal probabilities. However, sampling was not in general with replacement except in the case of the first stage units.

The simplified sampling scheme that we shall consider is as follows :

stage	number of units in		selection of sample is with replacement and probabilities
	population	sample	
first	$N$	$n$	$P_i$ for the $i$ -th first stage unit
second	$N_i$	$n_i$	$P_{ij}$ for the $j$ -th second stage unit in the $i$ -th first stage unit.
third	$N_{ij}$	$n_{ij}$	equal

Since the use of separate symbols for the variate value in the sample and in the population unnecessarily complicates the notation, we shall use the same symbol  $X_{ijk}$  to denote the variate value for the  $k$ -th third stage unit in the  $j$ -th second stage unit in the  $i$ -th first stage unit. The range of  $i, j, k$  will show whether we are referring to the sample or to the population. Furthermore, the dropping of a subscript will indicate a summation over the units in that stage, for instance,  $X_{ij}$  will stand for the total of the variate values for all third stage units in the  $j$ -th second stage unit of the  $i$ -th first stage unit, thus  $\sum_{k=1}^{N_{ij}} X_{ijk} = X_{ij}$ .

Similarly 
$$\sum_{j=1}^{N_i} X_{ij} = X_i \quad \text{and} \quad \sum_{i=1}^N X_i = X$$

We shall consider the problem of estimating the grand-total  $X$ .

From considerations of symmetry the following is taken as the estimate of  $X$

$$t = \frac{1}{n} \sum_{i=1}^n \frac{1}{P_i n_i} \sum_{j=1}^{n_i} \frac{N_{ij}}{P_{ij} n_{ij}} \sum_{k=1}^{n_{ij}} X_{ijk}$$

where, of course,  $X_{ijk}$  is the variate value for the  $k$ -th third stage unit in the  $j$ -th second-stage unit in the  $i$ -th first-stage unit in the selected sample. Obviously  $t$  is unbiased

$$E(t) = X$$

and a little computation shows that its variance is

$$V(t) = \frac{1}{n} \left\{ \sum_{i=1}^N \frac{\theta_i^2 + \sigma_i^2}{P_i n_i} + \sigma^2 \right\}$$

where

$$\sigma_{ij}^2 = \frac{1}{N_{ij}} \left\{ \sum_{k=1}^{N_{ij}} X_{ijk}^2 - \frac{X_{ij}^2}{N_{ij}} \right\}$$

$$\theta_i^2 = \sum_{j=1}^{N_i} \frac{N_{ij}^2}{P_{ij}} \frac{\sigma_{ij}^2}{n_i}$$

$$\sigma_i^2 = \sum_{j=1}^{N_i} \frac{X_{ij}^2}{P_{ij}} - N_i$$

$$\sigma^2 = \sum_{i=1}^N \frac{X_i^2}{P_i} - X^2$$

From the practical point of view, however, the numbers  $n_i$  and  $n_{ij}$  are not generally defined separately for each of the first and second stages. There are three different ways of fixing up the values of the  $n_i$ 's and  $n_{ij}$ 's. One we may call "equal sampling" at both stages where the same number  $m$  of second stage units are selected from each first stage unit and the same number  $l$  of third stage units are selected from within each selected



## VARIANCE COMPONENTS IN PROBABILITY SAMPLING

second stage unit. The other method may be called "proportionate sampling" at both stages where a fixed proportion of second stage units are selected within each first stage unit and a different fixed proportion of third stage units within each selected second stage unit are sampled. Variants of these two methods, with equal sampling at one stage and proportionate sampling at another stage are also used. A third method is to determine the values of  $n_i$  and  $n_{ij}$  in such a way that the estimate  $t$  comes out as proportional to the total of all variate values in the sample : this is known as "self-weighted" sampling.

Here we shall discuss two of these special types of sampling. The first is equal sampling at the second stage and proportionate sampling at the third. The other is self-weighted sampling with equal sampling at the second stage.

For equal sampling at the second stage and proportionate sampling at the third we have

$$n_i = m$$

$$n_{ij} = lN_{ij}$$

say. In this case,

$$t = \frac{1}{lmn} \sum_{i=1}^n \frac{1}{P_i} \sum_{j=1}^m \frac{1}{P_{ij}} \sum_{k=1}^{n_{ij}} X_{ijk}$$

and the variance reduces to

$$V(t) = \frac{A_1}{n} + \frac{B_1}{mn} + \frac{C_1}{lmn}$$

where  $A_1$ ,  $B_1$ ,  $C_1$  are independent of  $l$ ,  $m$ ,  $n$  and given by

$$A_1 = \sigma^2$$

$$B_1 = \sum_{i=1}^N \frac{\sigma_i^2}{P_i}$$

$$C_1 = \sum_{i=1}^N \frac{1}{P_i} \sum_{j=1}^{N_i} \frac{N_{ij} \sigma_{ij}^2}{P_{ij}}$$

We shall refer to this as the scheme I of sampling.

In the second scheme of sampling (Scheme II), we have

$$n_i = m$$

$$n_{ij} = \frac{l N_{ij}}{P_i P_{ij}}$$

so that the estimate comes out as

$$t = \frac{1}{lmn} \sum_{i=1}^n \sum_{j=1}^m \sum_{k=1}^{n_{ij}} X_{ijk}$$

with variance given by

$$V(t) = \frac{A_2}{n} + \frac{B_2}{mn} + \frac{C_2}{lmn}$$

where  $A_2, B_2, C_2$  are constants independent of  $l, m, n$  given by

$$A_2 = \sigma^2$$

$$B_2 = \sum_{i=1}^N \frac{\sigma_i^2}{P_i}$$

$$C_2 = \sum_{i=1}^N \sum_{j=1}^{N_i} N_{ij} \sigma_{ij}^2$$

so that  $A_1 = A_2, B_1 = B_2$  but  $C_1 \neq C_2$ . We shall write  $A$  for  $A_1$  or  $A_2$  and  $B$  for  $B_1$  or  $B_2$ .

### 3. THE STAGE COMPONENTS OF VARIANCE AND THEIR ESTIMATES

We thus see that in either of the two schemes of sampling under consideration the variance of the estimate depends on only three parameters  $A, B, C_i$  independent of the adjustable constants  $l, m, n$ . The cost of the survey naturally depends on the values of  $l, m, n$ . Therefore if estimates of the parameters  $A, B$ , and  $C_i$  are available, the information may be of use in planning an optimum survey at a fixed level of cost.

Let us now examine the nature of the three parameters. If it were possible to determine without further sampling the value  $X_i$  for a first stage unit selected with probabilities  $P_i$ ,  $\frac{X_i}{P_i}$  would provide an unbiased estimate of  $X$ . The parameter  $A$  simply measures the variance of such an estimate, that is  $A$  is the variance of the (hypothetical) estimate of  $X$  from complete enumeration of a single first stage unit chosen with the probabilities  $P_i$ . We may thus look upon  $A$  as the "between first stage" variance. Similarly  $\sigma_i^2$  gives the variance of the (hypothetical) estimate of  $X_i$  obtained by completely enumerating a second stage unit drawn with probabilities  $P_i$ . Thus  $\sigma_{ij}^2$  measures variation between second stage units within the  $i$ -th first stage unit. Therefore, if the  $N$  first stage units were regarded as strata and if from the  $i$ -th stratum  $\nu P_i$  second stage units were chosen with probabilities  $P_{ij}$  and each completely enumerated, the variance of the estimate of  $X$  would be simply  $B/\nu$ . The parameter  $C_i$  may be similarly interpreted this way.

We now take up the problem of estimating the parameters  $A, B, C_i$ . We shall simply obtain unbiased quadratic estimators for these parameters. Certain optimum properties of these estimators may be demonstrated from considerations of symmetry but we shall not enter here into a discussion of that type. For problems of estimation we shall consider unrestricted values of  $n_i$ 's and  $n_{ij}$ 's.

Let us write

$$s_{ij}^2 = \frac{1}{n_{ij}-1} \left( \sum_{k=1}^{n_{ij}} X_{ijk}^2 - \frac{x_{ij}^2}{n_{ij}} \right) \text{ where } x_{ij} = \sum_{k=1}^{n_{ij}} X_{ijk}$$

Then for fixed first and second stage units

$$E s_{ij}^2 = \sigma_{ij}^2$$

Hence, if we write

$$c_1 = \frac{1}{n} \sum_{i=1}^n \frac{1}{P_i^2 n_i} \sum_{j=1}^{n_i} \frac{N_{ij} s_{ij}^2}{P_{ij}^2}$$

$$c_2 = \frac{1}{n} \sum_{i=1}^n \frac{1}{P_i n_i} \sum_{j=1}^{n_i} \frac{N_{ij} s_{ij}^2}{P_{ij}}$$

then  $c_1$  and  $c_2$  provide unbiased estimators for  $C_1$  and  $C_2$  respectively.

Now construct

$$y_{ij} = \frac{1}{P_{ij}} \frac{N_{ij}}{n_{ij}} x_{ij}$$

Let

$$s_i^2 = \frac{1}{n_i - 1} \left( \sum_{j=1}^{n_i} y_{ij}^2 - \frac{y_i^2}{n_i} \right) \text{ where } y_i = \sum_{j=1}^{n_i} y_{ij}$$

For fixed first stage units  $E(s_i^2) = V(y_{ij})$ .

If first and second stage units are fixed

$$E y_{ij} = \frac{X_{ij}}{P_{ij}} \quad V(y_{ij}) = \frac{1}{P_{ij}^2} \frac{N_{ij}^2}{n_{ij}} \sigma_{ij}^2$$

and therefore when second stage units are allowed to vary, that is when only first stage units are fixed

$$\begin{aligned} V(y_{ij}) &= V \left( \frac{X_{ij}}{P_{ij}} \right) + E \left( \frac{1}{P_{ij}^2} \frac{N_{ij}^2}{n_{ij}} \sigma_{ij}^2 \right) \\ &= \sigma_i^2 + E \left( \frac{1}{P_{ij}^2} \frac{N_{ij}^2}{n_{ij}} \sigma_{ij}^2 \right) \end{aligned}$$

Therefore if we write

$$w_i^2 = s_i^2 - \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{1}{P_{ij}^2} \frac{N_{ij}^2}{n_{ij}} s_{ij}^2$$

for fixed first state unit  $E w_i^2 = \sigma_i^2$  and consequently

$$b = \frac{1}{n} \sum_{i=1}^n \frac{w_i^2}{P_i^2}$$

provides an unbiased estimate for  $B$ .

Finally to estimate  $A$  construct

$$z_i = \frac{1}{P_i n_i} y_i$$



Then, for fixed first stage unit,

$$E(z_i) = \frac{X_i}{P_i} \quad V(z_i) = \frac{1}{P_i^2 n_i} V(y_{ij})$$

and therefore for fixed first stage unit

$$E\left(\frac{s_i^2}{P_i^2 n_i}\right) = V(z_i)$$

Consequently for unrestricted variations

$$\begin{aligned} V(z_i) &= V\left(\frac{X_i}{P_i}\right) + E\{V(z_i)\} \\ &= \sigma^2 + E\{V(z_i)\}. \end{aligned}$$

Therefore, if we write

$$s^2 = \frac{1}{n-1} \left( \sum_{i=1}^n z_i^2 - \frac{z^2}{n} \right) \text{ where } z = \sum_{i=1}^n z_i$$

we have

$$a = s^2 - \frac{1}{n} \sum_{i=1}^n \frac{s_i^2}{p_i^2 n_i}$$

for an unbiased estimate of  $A$ .

We may note in this connection, the well known result that since

$$t = \frac{1}{n} \sum_{i=1}^n z_i$$

an unbiased estimate of its variance is given by  $\frac{s^2}{n}$  but if we are interested in the separate components of the variance, we have to compute  $a, b, c_i$  separately. One disadvantage of the estimates  $a, b, c_i$  is that sometimes these may turn out to be negative.

#### REFERENCES

- COCHRAN, W. G. (1939): The use of the analysis of variance in enumeration by sampling. *Amer. Stat. Ass.*, **34**, 492-510.
- LAHRI, D. B. (1954): Technical paper on some aspects of the development of the sample design. *The National Sample Survey* No. 5, Department of Economic Affairs, Ministry of Finance, Government of India.

Paper received : April, 1956.

# A NOTE ON TWO STAGE SAMPLING

By R. RANGARAJAN

Central Water and Power-Commission, New Delhi

## 1. INTRODUCTION

In a two stage sampling design, selection of the second stage units from primary units can be made in one of two ways. In the first place, the number of subunits to be selected from the primary ones may be fixed in advance so that whenever any primary unit is included in the sample, its subsampling size is known. The sample number hence will vary from sample to sample but a restriction on the variation may be placed by the condition that the expected value of sample number is a constant given in advance. Alternatively, the sample number may be fixed in advance and allocated among the primary units selected in a suitable manner. Godambe (1951) has examined the two situations for the case of the simple sampling scheme where the units in both stages were selected at random without replacements and has shown the gains in the efficiency of the estimate obtained in the former over that of the latter under optimum conditions. This note is intended to point out that this will be the case even in the most general type of sampling schemes with or without replacement in as much as the former lays a less stringent condition on the distribution of second stage sampling than the latter. The final results obtained in the two cases appear quite interesting.

## 2. SAMPLING WITHOUT REPLACEMENT

Let there be  $N$  primary units of sizes  $M_1, M_2, \dots, M_N$  from which a sample of  $n$  is chosen without replacement according to varying probabilities. Let  $p_i$  and  $p_{ij}$  be the probabilities respectively of the  $i$ -th p.s.u. and the  $i$ -th and the  $j$ -th p.s.u.'s to be included in the sample.

Also  $m_{si}$  ( $i = 1$  to  $n$ ) denote the sizes of secondary units chosen from the sample  $s$  of primary units ( $n$ ) with equal probability and without replacement.

Further,  $X_{ij}$  = the value of characteristic under study in the  $j$ -th subunit of  $i$ -th p.s.u.

$X_i$  = total over the subunits in the  $i$ -th p.s.u.

$X$  = total over all p.u.'s.

If the corresponding small letters represent sample values, the unbiased estimate of the total in the population and its variance can be given as

$$x_s = \sum_{i=1}^n \frac{M_i}{p_i} \sum_{j=1}^{m_{si}} \frac{x_{ij}}{m_{si}}, \quad \dots (2.1)$$

$$V(x_s) = \sum_{i=1}^N \frac{x_i^2(1-p_i)}{p_i} + \sum_{i \neq l=1}^N \frac{X_i X_l}{p_i p_l} (p_{il} - p_i p_l) +$$

$$+ \sum_s p(s/R) \left\{ \sum_{i=1}^n \frac{M_i^2}{p_i^2} \frac{M_i - m_i}{M_i m_i} \sigma_i^2 \right\} \quad \dots (2.2)$$

and

where 
$$\sigma_i^2 = \frac{\sum_{j=1}^{M_i} (X_{ij} - \bar{X}_i)^2}{M_i - 1}$$
,  $p(s/R)$  = probability of getting this particular set  $s$  of primary units selected and  $\sum$  denotes summation over all possible sets,  $(R)$ .

2.1. Now if  $m$ 's are fixed in advance, the optimum distribution of  $m_i$ 's is obtained by minimising the variance under condition

$$E(\sum m_{si}) = \sum_{i=1}^N p_i m_i = m_0 \quad (\text{given in advance}). \quad \text{Thus optimum}$$

$$m_i = \frac{m_0}{\left(\sum_{i=1}^N \frac{M_i \sigma_i}{p_i}\right)} \left(\frac{M_i \sigma_i}{p_i}\right).$$

$$\begin{aligned} \text{The minimum variance } V_E(x_s) &= I_1 + \sum_{i=1}^N \frac{M_i^2 \sigma_i^2}{p_i} \left( \frac{p_i}{m_0} \frac{\sum M_i \sigma_i}{M_i \sigma_i} - \frac{1}{M_i} \right) \\ &= I_1 + \frac{1}{m_0} \left( \sum_{i=1}^N M_i \sigma_i \right)^2 - \sum_{i=1}^N \frac{M_i \sigma_i^2}{p_i} \quad \dots (2.1.1) \end{aligned}$$

where  $I_1$  is the first two terms in (2.2).

2.2. If the  $m_i$ 's are chosen according to the sample selected, (the second case), we get

$$\sum_s m_i = m_0.$$

$$\text{Minimising condition gives, } m_{si} = \frac{m_0 \left( \frac{M_i \sigma_i}{p_i} \right)}{\left( \sum_{i=1}^n \frac{M_i \sigma_i}{p_i} \right)}.$$

$$\begin{aligned} \text{Minimum variance } V_P(x_s) &= I_1 + \sum_s p(s/R) \sum_{i=1}^n \frac{M_i^2 \sigma_i^2}{p_i^2} \cdot \left( \frac{\sum_{i=1}^n \frac{M_i \sigma_i}{p_i}}{m_0 \left( \frac{M_i \sigma_i}{p_i} \right)} - \frac{1}{M_i} \right) \\ &= I_1 - \sum_{i=1}^N \frac{M_i \sigma_i^2}{p_i} + \sum_s p(s/R) \left[ \frac{1}{m_0} \left( \sum_{i=1}^n \frac{M_i \sigma_i}{p_i} \right)^2 \right] \\ &= I_1 - \sum_{i=1}^N \frac{M_i \sigma_i^2}{p_i} + \frac{1}{m_0} \sum_{i=1}^N \frac{M_i^2 \sigma_i^2}{p_i} + \\ &\quad + \frac{1}{m_0} \sum_{i \neq k=1}^N \left( \frac{M_i \sigma_i}{p_i} \right) \left( \frac{M_k \sigma_k}{p_k} \right) p_{ik}. \quad \dots (2.2.1) \end{aligned}$$

$$\begin{aligned} V_P - V_E &= \frac{1}{m_0} \left[ \sum_{i=1}^N \frac{M_i^2 \sigma_i^2}{p_i} (1 - p_i) + \sum_{i \neq k=1}^N \left( \frac{M_i \sigma_i}{p_i} \right) \left( \frac{M_k \sigma_k}{p_k} \right) (p_{ik} - p_i p_k) \right] \\ &= \frac{1}{m_0} \text{var} \left( \sum_{i=1}^n \frac{M_i \sigma_i}{p_i} \right). \quad \dots (2.2.2) \end{aligned}$$



## A NOTE ON TWO STAGE SAMPLING

Thus the estimate in the first subsampling scheme is more efficient than that of the second one.

If the  $p_i$ 's are all equal, we immediately get the result derived by Godambe (1951).

### 3. SAMPLING WITH REPLACEMENT

In the case where the primary units are selected with replacement, using similar notations as above, the unbiased estimate of the total and its variance are given by

$$x_s = \frac{1}{n} \cdot \sum_{i=1}^n \frac{M_i \bar{x}_i}{p_i} \quad \dots (3.1)$$

where  $\bar{x}_i$  is the sample average of the  $i$ -th p.s.u. selected.

$$\begin{aligned} V(x_s) = & \frac{1}{n} \left[ \sum_{i=1}^N \frac{M_i^2 \bar{X}_i^2}{p_i} - \left( \sum_{i=1}^N M_i \bar{X}_i \right)^2 \right] + \\ & + \frac{1}{n^2} \cdot \sum_s p(s/R) \sum_{i=1}^n \frac{M_i^2}{p_i^2} \cdot \frac{M_i - m_i}{M_i m_i} \sigma_i^2 \quad \dots (3.2) \end{aligned}$$

where  $\sigma_i^2$ ,  $p(s/R)$ ,  $\sum_s$  are as defined earlier.

3.1. By the first mode of subsampling, we have

$$E_s \left( \sum_{i=1}^n m_{si} \right) = n \sum_{i=1}^N p_i m_i = m_0,$$

$$\text{optimum } m_i = \frac{m_0}{n \left( \sum_{i=1}^N M_i \sigma_i \right)} \left( \frac{M_i \sigma_i}{p_i} \right).$$

$$\text{Minimum variance } V_E(x_s) = I_1 + \frac{1}{m_0} \cdot \left( \sum_{i=1}^N M_i \sigma_i \right)^2 - \frac{1}{n} \sum_{i=1}^N \frac{M_i \sigma_i^2}{p_i} \quad \dots (3.1.1)$$

3.2. According to the second method,

$$\sum_{i=1}^n m_i = m_0,$$

$$\text{which gives optimum } m_i = m_0 \cdot \left( \frac{M_i \sigma_i}{p_i} \right) / \sum_{i=1}^n \frac{M_i \sigma_i}{p_i}.$$

$$\begin{aligned}
 \text{Minimum variance } V_F(x_s) &= I_1 + \sum_s p(s/R) \sum_{i=1}^n \frac{M_i^2 \sigma_i^2}{n^2 p_i^2} \left\{ \frac{\sum_{i=1}^n \frac{M_i \sigma_i}{p_i}}{m_0 \left( \frac{M_i \sigma_i}{p_i} \right)} - \frac{1}{M_i} \right\} \\
 &= I_1 + \sum_s \frac{p(s/R)}{m_0 n^2} \left( \sum_{i=1}^n \frac{M_i \sigma_i}{p_i} \right)^2 - \sum_s \frac{p(s/R)}{n^2} \cdot \sum_{i=1}^n \frac{M_i \sigma_i^2}{p_i^2} \\
 &= I_1 + \frac{1}{m_0 n} \cdot \sum_{i=1}^N \frac{M_i^2 \sigma_i^2}{p_i} + \frac{n-1}{m_0 n} \cdot \left( \sum_{i=1}^N M_i \sigma_i \right)^2 - \frac{1}{n} \cdot \sum_{i=1}^N \frac{M_i \sigma_i^2}{p_i} \\
 &\quad \dots (3.2.1)
 \end{aligned}$$

$$\begin{aligned}
 V_F - V_E &= \frac{1}{m_0 n} \cdot \left[ \sum_{i=1}^N \frac{M_i^2 \sigma_i^2}{p_i} - \left( \sum_{i=1}^N M_i \sigma_i \right)^2 \right] \\
 &= \frac{1}{m_0} \cdot \text{var} \left( \frac{1}{n} \cdot \sum_{i=1}^n \frac{M_i \sigma_i}{p_i} \right). \quad \dots (3.2.2)
 \end{aligned}$$

It is interesting to note that the difference between the variance of the estimates obtained in the two modes of subsampling is exactly equal to  $\frac{1}{m_0}$  times the variance of the unbiased estimate of  $\left( \sum_{i=1}^N M_i \sigma_i \right)$  from a single stage sampling with the corresponding varying probabilities, with and without replacements respectively. Further the second mode of subsampling which is usually adopted in practice is always less efficient than the first.

I am thankful to Dr. Des Raj for suggesting the problem.

#### REFERENCES

- COCHRAN, W. G. (1953): *Sampling Techniques*, Wiley & Sons, New York.  
 GODAMBE (1951): On two stage sampling. *J. Roy. Stat. Soc.*, B, 13, 216—218.  
 HORVITZ, D. G. AND THOMPSON D. J. (1952): A generalisation of sampling without replacement from finite universe. *J. Amer. Stat. Ass.*, 47, 663-685.

*Paper received: June, 1956.*

# METHOD OF MATCHING USED FOR THE ESTIMATION OF TEST RELIABILITY \*

By P. K. BOSE and S. B. CHAUDHURI  
*University of Calcutta*

## 1. INTRODUCTION

Reliability of a test has been defined as the correlation between two parallel tests, i.e., tests with equal means, equal variances and equal intercorrelations which, however, was found to be the same as the ratio of true variance to the observed variance of the scores in test under the additive set up i.e. observed score is the sum of true score and an error component.

Different methods have been developed to estimate the reliability of a test.

Three important methods for estimating test reliability are

- (1) method of parallel forms,
- (2) test-retest method,
- (3) split-half method.

Leaving aside the detailed discussion of these methods it may only be pointed out that the first two methods do not very often find their application. As to the first one the difficulty lies with obtaining parallel forms of a test and also with applying the different forms on the same group at different occasions. Similarly regarding the second one it becomes difficult to retest the same group of individuals and even if the group be available, the question of time interval between the two tests comes in. Long gap and short gap both give rise to serious bias in the estimation.

Because of these difficulties we are to consider the possibility of obtaining an estimate of the reliability from only one form of a test and from only one application of it over a group of individuals. This can be done by the third method referred to above, i.e., the split-half method according to which the test is to be subdivided into two equal subgroups each containing an equal number of items of the original test. The correlation coefficient between the two subtests gives an estimate of the reliability of each half. Then Spearman-Brown's formula may be applied to obtain the reliability of the whole test.

Various methods of splitting the test have been suggested, namely, first versus second halves, odd versus even items, etc. But these methods of splitting do not appear to be appropriate because according to the very definition of reliability the two subtests should be parallel which is also necessary for the application of Spearman-Brown's formula. But according to the above methods of splitting this criterion is seldom satisfied. On the other hand there are  $n!/2 \left(\frac{n}{2}!\right)^2$  different ways of dividing a test of  $n$  items into two halves leading to  $n!/2 \left(\frac{n}{2}!\right)^2$  different estimates of reliability. All these ways of splitting are not equally

---

\* Read at the Statistics Section of the Indian Science Congress, 1956, at Agra.



defendable and the different estimates largely fluctuate from one another. This defect has been referred to by Kuder and Richardson (1952). In this paper we shall describe a method of splitting a test by matching the test items on the basis of item analysis. This will satisfy the criterion referred to above and will give a unique value to the estimate.

## 2. METHOD

Let us define the difficulty value of an item as the proportion of individuals who are unsuccessful in answering that item, non-attempt being considered equivalent to failure. Let us consider a sub-test of  $k$  items the score for the  $i$ -th item being  $x_i$  ( $i = 1, 2, \dots, k$ ) which takes the value 1 for pass and 0 for failure. Also let the difficulty value of the  $i$ -th item be  $q_i$  ( $i = 1, 2, \dots, k$ ). Then we have

$$\begin{aligned} \text{mean of} \quad & x_i = 1 - q_i = p_i, \\ \text{variance of} \quad & x_i = q_i(1 - q_i) = p_i q_i \\ & (i = 1, 2, \dots, k). \end{aligned}$$

Then the mean and variance of the sub-test will be obtained as follows.

$$\begin{aligned} \text{Mean of the sub-test score} &= \text{mean} \left( \sum_{i=1}^k x_i \right) \\ &= \sum_{i=1}^k \text{mean} (x_i) \\ &= \sum_{i=1}^k p_i. \end{aligned} \quad \dots (2.1)$$

$$\begin{aligned} \text{Variance of the sub-test score} &= \text{variance} \left( \sum_{i=1}^k x_i \right) \\ &= \sum_{i=1}^k p_i q_i + \sum_{i \neq j}^k r_{ij} \sqrt{p_i q_i \cdot p_j q_j} \end{aligned} \quad \dots (2.2)$$

where  $r_{ij}$  is the correlation between the  $i$ -th and the  $j$ -th items in the subtest.

We define the two subtests to be properly matched when the difficulty values of the items of one are equal to the difficulty values of the corresponding items of the other. Thus if we form the two subtests such that the difficulty values of the corresponding items are the same, then from (2.1) and (2.2) we find that the mean of the two subtests are equal and the variance of the two subtests are equal provided that the correlation coefficient between any pair of items remains the same as long as the difficulty values of the items in the pair do not change, i.e., the correlation coefficient is a function of difficulty values only. This is a mild restriction and is satisfied in almost all cases. Thus if we match the different items in the two subtests in this way, the two subtests become parallel.

Split-half method is only a particular case of split-multiple method according to which the whole test is subdivided into a number of subtests each containing the same number of items. Let us first obtain an expression for the covariance between two such subtests. Let each of them contain  $k$  items, the score for the  $i$ -th item being  $x_i$  and  $x'_i$

respectively ( $i = 1, 2, \dots, k$ ) and difficulty values being  $q_i$  and  $q'_i$  respectively. Then we have covariance between these two subtest scores

$$\begin{aligned} &= \text{cov} \left\{ \sum_{i=1}^k x_i, \sum_{j=1}^k x'_j \right\} \\ &= \sum_{i,j=1}^k \text{cov} (x_i, x'_j) \\ &= \sum_{i,j=1}^k r'_{ij} \sqrt{p_i q_i \cdot p'_j q'_j} \quad \dots (2.3) \end{aligned}$$

where  $r'_{ij}$  is the correlation between the  $i$ -th item of one and  $j$ -th item of the other subtest.

Then if we form the different subtests such that the difficulty values of the items of one are equal to the difficulty values of the items of any one else the subtests will be parallel in the sense that they have equal means, equal variances and equal inter-correlations by relations (2.1), (2.2) and (2.3) under the mild restriction stated above in connection with split-half technique. The advantage of split-multiple technique (with at least 3 parts) over the split-half is that in the former it is possible to make a more complete check by Wilks' (1946)  $L_{mve}$  test and it is possible to be sure that the forms are parallel, not only with respect to means and variances but also with respect to correlations.

It may be pointed out that when in the population the split-multiple test scores have equal means, equal variances and equal inter-correlations, sample observations may not exhibit similar values although the values may not be significantly different. In the case of non-significance only reliability may be estimated. In the case of split-half one gets unique value of correlation between the two halves; but that is not so, for split-multiple case — there will be a number of inter-correlations between two parallel subtests. This can be made unique by utilising the pooled estimate—this can be taken as a satisfactory estimate of the population value.

Thus far we have described the method of splitting a test on the basis of difficulty values of the items. But when this type of splitting is not possible in view of the fact that we do not get  $m$  sets of items ( $m \geq 2$ ) the difficulty values of one being equal to those of other we split up the test on the basis of mean difficulty values — i.e. subtests are so formed that the mean difficulty value for each remains the same so that the subtests are parallel at least with respect to means.

Another point that must be noted is that the difficulty values are generally calculated on the basis of sample observations. Thus in the case of small differences in difficulty values of several items we are to test whether the difference is real or may be accounted for by sampling fluctuations. As some amount of correlation is always present between any two test items the proper test for matching in this case will be Cochran's  $Q$  test (1950), a short description of which is given below.

Cochran has extended the familiar  $\chi^2$ -test for comparing the percentages of successes, in a number of independent samples to the situation in which each member of any sample is matched in some way with a member of every other sample so that some correlation is introduced between the results in different samples. In the present case the same individual is selected as a sample member in connection with all the test items; thus it is a case of strictly matched sample. For the case of 2 samples an appropriate test was easily constructed by

McNemar. But Cochran's test is more general and can be used for 2 or more samples. According to this test procedure the data are arranged in a two-way table with  $r$  rows and  $c$  columns in which each column represents a sample and each row a matched group.

The test criterion is

$$Q = \frac{c(c-1)\sum_j (T_j - \bar{T})^2}{c \sum_i u_i - (\sum_i u_i^2)} \quad \dots (2.4)$$

where

$T_j$  = total number of successes in the  $j$ -th sample (column)  
 $u_i$  = total number of successes in the  $i$ -th row.

If the true probability of success be the same in all samples and if the number of rows be large this  $Q$  in the limiting case follows  $\chi^2$ -distribution with  $(c-1)$  d.f. In case there are only 2 samples this  $Q$  follows strictly the  $\chi^2$ -distribution with 1 d.f. In this case the expression for  $Q$  may be simplified to

$$Q = \frac{(b-c)^2}{(b+c)} \quad \dots (2.5)$$

where the corresponding 2-way table will be as shown below:

TABLE 1. THE TWO WAY TABLE

sample I \ sample II	success	failure
success	$a$	$b$
failure	$c$	$d$

Actually McNemar (1949) got his result in this form. In this form the only things to be known are the values of  $b$  and  $c$ , i.e., the number of matched groups having success in the first and failure in the second and having failure in the first and success in the second.

After forming the parallel subtests one can calculate the correlation between two such subtests. The maximum likelihood estimate of  $\rho$  obtained from a sample from a normal population specified by three parameters  $\sigma_x = \sigma_y$ ,  $\mu_x = \mu_y$ , and  $\rho$  is

$$\rho = \frac{2\sum XY - \frac{(\sum X + \sum Y)^2}{2N}}{\sum X^2 + \sum Y^2 - \frac{(\sum X + \sum Y)^2}{2N}} \quad \dots (2.6)$$

This should be the appropriate formula which should be used for the estimation of test reliability of the subtest. We can then apply Spearman-Brown's formula for multiple length to get the reliability for the whole test

$$R_{kk} = \frac{kR_{11}}{1 + (k-1)R_{11}} \quad \dots (2.7)$$



## METHOD OF MATCHING FOR ESTIMATION OF TEST RELIABILITY

where  $R_{11}$  = reliability of a test of unit length,  
and  $R_{kk}$  = reliability of a test of length  $k$  units.

When the whole test cannot be divided into a number of subtests of equal length in view of the fact that  $n$  (the number of items in the whole test) is not an exact multiple of 2, 3, 4, 5 etc., we can form some subtests of equal length and only one of lesser length. Let us call the subtests of equal length as full and that of lesser length as fractional. We may calculate the reliability of one full subtest and may then apply Spearman-Brown's formula for multiple length. This multiple is, however, not integral.

### 3. COMPARISON WITH KUDER AND RICHARDSON'S FORMULA

In Kuder and Richardson's well-known formula for reliability coefficient

$$r_{tt} = \frac{n}{n-1} \cdot \frac{\sigma_t^2 - \sum_{i=1}^n p_i q_i}{\sigma_t^2} \quad \dots (3.1)$$

where  $n$  stands for the number of items in the test and  $\sigma_t^2$  stands for the variance of the test, two assumptions were made, namely, (i) the items have equal standard deviations and (ii) the inter-item correlations are equal. In the method which has been discussed no such assumptions are necessary. The only assumption that has been made is that intercorrelations of pairs of test items having the same pair of difficulty values are the same. For the purpose of splitting we should only guarantee that the items are properly matched with respect to their difficulty values. So far as the time and difficulty of computations are concerned the split multiple method as described above and the Kuder-Richardson's method stand on the same level.

### 4. AN ILLUSTRATION : RELIABILITY COEFFICIENT ESTIMATED

#### (i) *Method of matched items.*

Scores of a random sample of 500 candidates have been obtained from the records of a scholastic test. Each candidate has got six scores, i.e., scores in six different items of the test, namely question nos. 1, 2, 3, 4, 5 and 6. The object of the investigation is to find out the reliability of the whole test after splitting it up into two halves.

The original scores are translated to a new system namely for each individual and for each item a score of one is given in the case of pass and zero in the case of failure.

From the total number of passes in the different test items, difficulty values were calculated. These were found to be as follows:

TABLE 2. DIFFICULTY VALUES OF TEST ITEMS

test item number	number of passes	$= 1 - \frac{\text{difficulty value}}{\text{total no. of pass}}$ $\frac{\text{total no. of pass}}{\text{total no. of candidates}}$
I	427	0.146
II	448	0.104
III	420	0.160
IV	426	0.148
V	442	0.116
VI	435	0.130

Then  $Q$  test was applied to find out the test items which may be considered as equivalent in the sense that the corresponding population difficulty values are identical. We try to form pairs containing two equivalent tests.

The two-way table and the  $Q$ -values are given below.

TABLE 3. THE TWO WAY TABLE ( $Q=.333$ )

test item IV test item III	success	failure
success	369	51
failure	57	23

TABLE 4. THE TWO WAY TABLE ( $Q=.561$ )

test item VI test item I	success	failure
success	384	53
failure	61	12

TABLE 5. THE TWO WAY TABLE ( $Q=.400$ )

test item II test item V.	success	failure
success	400	42
failure	48	10

All these  $Q (= \chi^2)$  values are non-significant at 5% level of significance, the 5% value being 3.841. We may now group the items as follows:

subtest A                      test items  
subtest B                      I, II, and IV.  
   III, V, and VI.

We then get the following results

mean (A) = 2.602,      var (A) = .443596,  
mean (B) = 2.594,      var (B) = .493164,      cov (A, B) = .158412

Then by using the formula (2.6),

$$\text{cor (A, B)} = .338167.$$

Thus the reliability of each subtest may be taken as .338167. By Spearman-Brown's formula the reliability of the whole test will become

$$\frac{2 \times .338167}{1 + .338167} = .505418.$$

## METHOD OF MATCHING FOR ESTIMATION OF TEST RELIABILITY

Next we try to form three subtests each containing two test items. For this purpose we are to find out two sets of 3 equivalent tests items by the application of  $Q$  test. The three subtests may be taken as

	test items
subtest A :	II and III
subtest B :	IV and V
subtest C :	I and VI

We then get following results:

mean (A) = 1.736	var (A) = 0.274304	cov (A, B) = 0.062304
mean (B) = 1.736	var (B) = 0.270304	cov (A, C) = 0.079136
mean (C) = 1.724	var (C) = 0.247824	cov (B, C) = 0.077136
	cor (A, B) = 0.228803	
	cor (A, C) = 0.302949	
	cor (B, C) = 0.297569.	

Thus the reliability of each subtest may be taken as the pooled correlation coefficient i.e. 0.276440.

By Spearman-Brown's formula the reliability of the whole test will become

$$\frac{3 \times 0.276440}{1 + 2 \times 0.276440} = .534053.$$

(ii) *Kuder and Richardson's method.*

TABLE 6.  $p_i$  AND  $q_i$  VALUES

test item number	number of passes	$q_i$	$p_i$
I	427	.146	.854
II	448	.104	.896
III	420	.160	.840
IV	426	.148	.852
V	442	.116	.884
VI	435	.130	.870

The above table gives  $p_i$  and  $q_i$  values for the different test items. The variance of the total test has been found out to be 1.253584. Then by using the formula

$$\text{reliability} = \frac{6}{5} \frac{1.253584 - .694008}{1.253584} = .535657$$



## CONCLUSION

A method of matching has been proposed for the estimation of test reliability. The two methods (namely the method of matching and Kuder-Richardson's method) give quite close estimates for the reliability coefficient.

## REFERENCES

- COCHRAN, W. G. (1950): The comparison of percentages in matched samples. *Biometrika*, **37**, 256-266.
- KUDER, G. F., AND RICHARDSON, M. W. (1937): The theory of the estimation of test reliability. *Psychometrika*, **2**, 151-160.
- MCNEMAR, Q. (1949): *The Psychological Statistics*, John Wiley & Sons., New York.
- WILKS, S. S. (1946): Sample criteria for testing equality of means, equality of variances and equality of covariances in a normal multivariate distribution. *Ann. Math. Stat.*, **17**, 257-281.

*Paper received : March, 1956.*

# RECOMMENDATIONS FOR PERSONNEL SELECTION IN INDIA BASED ON THE BRITISH SELECTION METHODS IN THE CIVIL SERVICE AND INDUSTRY<sup>1</sup>

By RHEA S. DAS

*Indian Statistical Institute, Calcutta*

## SUMMARY

Selection procedures of the Civil Service and industry in Britain are reviewed. Two main types of procedure, paper and pencil tests and selection boards, are described in some detail, and validity data are presented where available. Critical observations on these procedures, their use and implications, are also detailed. The possible application, and suggested areas of modification, of these procedures for use in India follows the British survey. Some current work utilizing these procedures is briefly indicated.

## PERSONNEL SELECTION METHODS

This report is based on interviews with leading British psychologists and on a survey of the research literature published from 1950 to 1956.

In Great Britain, as elsewhere, the primary objective of personnel selection is to predict from a field of applicants for a position, those applicants who will be most successful in the position. The task of prediction is qualified by considerations of economy, both of time and money. It is necessary, therefore, to devise methods for obtaining an optimum amount of information with minimum time and cost.

There are two types of personnel selection methods currently in use in Britain, both in the Civil Service and in industry :

- 1) paper and pencil tests of abilities,
- 2) selection boards.

In comprehensive selection programmes, as for the Administrative Class of the Civil Service, both of these types are utilized. Depending on the nature of the job, however, usually one or the other type will be chosen. Considerations relevant to the choice of type will be presented later in this report.

*Paper and Pencil tests :* Paper and pencil tests of ability have found wide use in the Civil Service and have also been used extensively in industrial selection by the National Institute of Industrial Psychology (NIIP). The design and content of several batteries of such tests will illustrate their general make-up. The basic test battery of the Civil Service for the Administrative Class selection consists of the following tests : NIIP test 70/1 for non-verbal intelligence; a verbal facility test; Babington Smith's advanced verbal intelligence test; and various versions of a test of general information, dealing chiefly with current affairs (Vernon, 1950). At other levels of the Civil Service, non-administrative in nature, paper and pencil tests are also used. The general pattern of these tests only can be given for security reasons. A specific test battery, consisting of a variety of tests, has been developed for each of the different grades of the Civil Service. While some tests are similar for all grades, others are specific to a particular grade. Those tests which are similar in all grades include verbal and non-verbal intelligence and arithmetic, and differ only in level of complexity with respect

---

<sup>1</sup> This report is based on a survey conducted in Great Britain by the author as a Research Fellow of the Indian Statistical Institute.

to the grade. The verbal intelligence tests include verbal analogies, grammar and verbal comprehension; the non-verbal intelligence tests include pictorial or diagrammatic materials presented in problem form; the arithmetic tests are of the customary type. Examples of tests specific for a particular grade include the clerical tests for the clerical grades and geography tests for Post Office Counter Clerks. Further tests used in the Ministry of Labour have been developed by psychologists at Birkbeck College in London. The content of these tests cannot be published for security reasons, however the tests have been modelled after the Admiralty Tests, described by Vernon and Parry (1949). The Admiralty battery consisted of the Shipley abstractions (Shipley and Burlingame, 1941), the Bennett Mechanical Comprehension Test (Bennett, 1948), Raven's Progressive Matrices (Raven, 1938), the NIIP Squares Test, a simple arithmetic test, a test of mathematical knowledge, and a test of mechanical and electrical knowledge. The last three tests were devised by the Admiralty psychologists.

Several different test batteries have been developed for and used by industry. The factory operatives selection battery for the Rowntrees Cocoa Works includes the following tests: an English achievement test; an arithmetic achievement test; intelligence tests of the abstraction type; and performance tests designed for use in the factory (Porteus, 1950). Use of NIIP tests in industrial selection is discussed by Castle and Garforth (1951) and Handyside and Duncan (1954). NIIP has developed the following tests for use in industrial selection and vocational guidance: Group Test (GT) 20, "Accuracy in checking names and numbers"; GT 25, "Aptitude for general clerical work"; GT 33, 36, and 90A, verbal intelligence tests for different age groups; GT 70/1 and 70/23, non-verbal intelligence tests; GT 81, Squares, and Form Relations, space perception tests; and the Vincent Models Tests, a mechanical aptitude test.

Published validities for the Civil Service and industrial tests, the nature of the validation criterion and test population, and the significance of the coefficients not corrected for restriction of range are summarized in Table 1.

TABLE 1. VALIDITY OF SELECTION TESTS

test name	N	r	P	$r^1$	nature of criterion	nature of sample	reference
NIIP GT 70/1	202	-.069	—	-.042	supervisor's ratings two years after selection	administrative class, civil service	Vernon (1950)
verbal facility	149	.172	.05	.223	supervisor's ratings two years after selection	administrative class, civil service	Vernon (1950)
B. Smith advanced verbal intelligence.	150	.083	—	.240	supervisor's ratings two years after selection	administrative class, civil service	Vernon (1950)
NIIP GT 33 & 70/23	44	.59	.01	—	manager's ratings two years after selection	factory supervisors	Castle and Garforth (1951)
NIIP GT 33 & 70/23	44	.51	.01	—	composite criterion <sup>2</sup> four years after selection	factory supervisors	Handyside and Duncan (1954)
general information	202	.030	—	-.149	supervisor's ratings two years after selection	administrative class, civil service	Vernon (1950)

<sup>1</sup> Corrected for restriction of range by Vernon (1950).

<sup>2</sup> Composite criterion based on manager's assessments and incidence of promotion.



## SELECTION METHODS IN THE CIVIL SERVICE AND INDUSTRY

The reported correlations for the Civil Service Administrative Class sample are consistently low. These values may be attributed to two factors : first, the highly homogeneous nature of the sample with respect to intellectual performance; and second, the collection of validation data on the successful candidates solely. The effect of these two factors would be to restrict variability, and hence to reduce the correlation coefficient. The higher correlations for the industrial samples reported in Table 1 may be due to the greater heterogeneity of the test sample with respect to the tested characteristics. Readers interested in the corrections for restricted variability, or restriction of range, are referred to Gulliksen (1950).

On a theoretical level, the tests described above adhere to Spearman's concept of "g" or general intelligence, which means that intelligence is treated as a unitary phenomenon rather than consisting of multiple abilities (Anastasi, 1954). In particular this is true of the Civil Service intelligence tests and the NIIP intelligence tests used in industrial selection. At the operational level, the addition of other tests to the selection batteries, such as mechanical comprehension and clerical aptitude, indicates acceptance of intelligence as consisting of multiple abilities. This contradiction that exists between the theoretical and operational levels should be resolved by empirical research on tests in relation to job performance. The resulting information would possibly enhance the predictive value of the selection tests in the Civil Service and in industry.

Another general characteristic of the test batteries described above is that they show little departure from the common pattern, i.e., verbal and nonverbal tests of intelligence, arithmetic tests, and several specific ability tests. In the field of selection, it may also be noted that there have been relatively few published reports of the validity of the various tests. Finally, it should be observed that the majority of the selection material used in Britain comes from three sources : the British forces (Vernon and Parry, 1949); the Civil Service (Wilson, 1948; Vernon, 1950); and NIIP (Castle and Garforth, 1951).

*Selection Boards :* The second major type of selection procedure, selection boards, was pioneered in Britain by the military forces' War Office Selection Boards (Vernon and Parry, 1949). Selection boards were later adopted by the Civil Service and NIIP. This method is generally known as a "new-type selection board" or Civil Service Selection Board (CISSB). The selection board can be most completely illustrated by reference to the CISSB procedure (Civil Service Commissioners, 1951; Vernon, 1950; Wilson, 1948).

"CISSB" is actually the second stage in Method II of the Normal Open Competitions for the Administrative Class and Foreign Service of the British Civil Service (Civil Service Commissioners, 1951), and consists of a series of tests and interviews requiring two to three days. The tests are divided into two classes, "psychological tests" and "analogous tests". The "psychological tests" include tests of ability and personality, with the ability tests stressing verbal and reasoning facility, and the personality tests stressing motivation, interests and personal history. "Analogous tests" are designed to be comparable to situations which successful candidates will have to face in the Civil Service. Two of these tests are based upon a lengthy dossier describing an imaginary problem, e.g., setting up an atomic reactor station in a hypothetical town, or government sponsored emigration from some area in England to Australia. In the first analogous test candidates are required to write on a question of policy or principle after careful study of the dossier. In the second analogous test, every candidate is allotted a special policy problem related to the dossier, which he as

chairman must present to the committee (the other candidates) and lead the committee to a solution of the problem. Oral and written exercises on general problems relevant to the Civil Servant complete the analogous tests. Three interviews in addition to the written and oral tests complete the schedule of the CISSB board. The first interview is a viva or oral examination of intellectual performance, conducted by one of the staff members; the second interview is a personality oriented interview, conducted by the psychologist; and the third interview is a general interview in the traditional manner by a third member of the staff. These three staff members, making up the Directing Staff, assess each candidate in the written and oral analogous exercises and in interview.

Industrial use of CISSB type selection boards is reported by NIIP investigators (Fraser, 1950; Castle and Garforth, 1951) and the industrial firm "Rowntrees" (Higham, 1952). In these reports, the selection boards are referred to as "new-type" selection boards. Between 1945 and 1950, about fifty of these selection boards had been carried out for industrial appointments by NIIP (Fraser, 1950). The majority of these selection boards were concerned with managerial positions or management trainees, but some were also concerned with sales, supervisory, and professional appointments. The NIIP procedure includes tests, interviews, and analogous exercises (Castle and Garforth, 1951; Fraser, 1950; Handyside and Duncan, 1954). Choice of paper and pencil tests was usually made from the NIIP collection, previously listed. Two interviews were included in the board procedure, one by the psychologist, and one by the entire board staff. Analogous exercises utilized concrete industrial problems which were similar to those successful candidates would face in the job. NIIP considers the psychologist's interview to be the focal point of the selection board: it outlines the main behaviour patterns, while the analogous tests confirm these patterns, and the paper and pencil tests indicate the limits to which a candidate may be expected to develop (Fraser, 1950).

Industrial application of selection boards is also illustrated by Rowntrees Cocoa Works (Higham, 1952). Tests, interviews, and analogous problems again make up the board schedule, which is generally used in selection of salesmen, although it is also used for depot managers, trainee overlookers, and other posts. Analogous problems, performed in a group situation, provide information on the ability of candidates to handle ideas, their effectiveness as group members, and their reactions to stress. These analogous problems are designed to be complex in nature and to admit of various solutions, and are discussed in a general group discussion as well as committee type situation with each candidate acting as chairman in turn.

Three components in the selection boards are found in common in the above reports: the use of tests of ability and other characteristics; two or more interviews; and exercises analogous to the job task, partially or completely performed in a group situation. Validities for the test, interview, and analogous exercise components of the selection board, and the validity of the total selection board, are reported by some investigators, and are summarized in Table 2. The significance of the validity coefficients, nature of the criterion and test population are also summarized in Table 2.

Review of Table 2 shows relatively higher validities than those reported in Table 1. It is possible that this general difference is due to the pooling of more than one test in Table 2 having the effect of increasing reliability which in turn increases the validity coefficient. All the coefficients reported in Table 2 differed significantly from zero at the .01 point of confidence.



# SELECTION METHODS IN THE CIVIL SERVICE AND INDUSTRY

TABLE 2. VALIDITY OF SELECTION BOARDS

procedure	<i>N</i>	<i>r</i>	<i>P</i>	nature of criterion	nature of sample	reference
tests	202	.315	.01	supervisor's ratings two years after selection	administrative class, civil service	Vernon (1950)
	44	.59	.01	manager's ratings two years after selection	factory supervisors	Castle and Garforth (1951)
	44	.51	.01	composite criterion <sup>1</sup> four years after selection	factory supervisors	Handyside and Duncan (1954)
interviews	202	.487	.01	supervisor's ratings two years after selection	administrative class, civil service	Vernon (1950)
	44	.66	.01	manager's ratings two years after selection	factory supervisors	Castle and Garforth (1951)
	44	.55	.01	composite criterion <sup>1</sup> four years after selection	factory supervisors	Handyside and Duncan (1954)
analogous tests	202	.445	.01	supervisor's ratings two years after selection	administrative class, civil service	Vernon (1950)
	44	.63	.01	manager's ratings two years after selection	factory supervisors	Castle and Garforth (1951)
	28	.58 <sup>2</sup>	— <sup>3</sup>	composite criterion four years after selection	factory supervisors	Handyside and Duncan (1954)
total procedure	202	.505	.01	supervisor's ratings two years after selection	administrative class, civil service	Vernon (1950)
	44	.68	.01	manager's ratings two years after selection	factory supervisors	Castle and Garforth (1951)
	44	.65	.01	composite criterion <sup>1</sup> four years after selection	factory supervisors	Handyside and Duncan (1954)

<sup>1</sup> Composite criterion made up of manager's ratings and incidence of promotion.

<sup>2</sup> Corrected for selectivity.

<sup>3</sup> Uncorrected coefficient not given for determining significance.



The relative value of the coefficients reported in Table 2 is made more meaningful in terms of selection procedures by reference to Table 3, which summarizes data from Castle and Garforth (1951) and Handyside and Duncan (1954). In these two studies, systematic or experimental selection was compared with the customary or control method of selecting supervisors in a large heavy engineering firm in Scotland. All men selected by either procedure were advanced to supervisory posts, permitting later comparison of the two methods. The superior prediction of the experimental procedure over the control procedure, in terms of the correlation with the criteria, is evident.

TABLE 3. EXPERIMENTAL COMPARISON OF SELECTION PROCEDURES

criterion	N	correlation of criterion with		reference
		control procedure	experimental procedure	
manager's ratings two years after selection	44	.23	.68	Castle and Garforth (1951)
composite criterion <sup>1</sup> four years after selection	44	.18	.65	Handyside and Duncan (1954)

<sup>1</sup> Composite criterion derived from manager's ratings and incidence of promotion.

Recent work in British personnel selection has been reviewed above, in terms of major types of procedures, and with special attention to validity data. It is now pertinent to consider how and where such procedures may be applied in India. The following points are to be discussed :

- 1) areas for which each type of procedure is suited;
- 2) modifications desirable for their application in India;
- 3) current use of such techniques.

It will be recalled that in the British Civil Service, only paper and pencil tests were used for selection of non-administrative grades, while selection boards were limited to selection of administrative grades. It is presumed that the basic reason for this practice lies in economy, both of time and money. The selection boards are more expensive in terms of these factors than the paper and pencil tests alone, and hence must be used only where this greater cost is justified. The assumption of CISSB is that a person who ultimately fails in his job at the administrative level is more costly than the insurance money spent in the more detailed and elaborate selection procedures. Hence, considerations of long run economy are operating even in this case. In the case of the non-administrative classes, the greater replaceability of failures leads to the choice of less expensive selection procedures. These considerations would appear to be applicable also in India. For the majority of posts, especially with a job scarcity and large labour market, the more economical procedures would be most practical. With higher level posts, of administrative, policy-making character, more detailed selection methods may be desirable. Where factors of group cooperation, personal contact, and leadership are important, as in the village development schemes, the more elaborate selection boards may also be of value.

The question next arises as to how these procedures should be modified to suit Indian conditions. With respect to paper and pencil tests, and tests used in selection boards,

## SELECTION METHODS IN THE CIVIL SERVICE AND INDUSTRY

validation of the tests in India is of primary importance. Language factors and items with cultural connotations should be altered, where foreign tests are used, to be more understandable to the test population. In addition to these factors, investigators in India may benefit by reviewing British efforts, and develop efficient batteries of tests which agree at both theoretical and empirical levels. These batteries should be based on thorough job analysis, which will improve prediction of persons who will ultimately be successful in the job (Stevenson, 1951). Interview procedures, already in wide use, do not require modification in terms of cultural context: they may benefit, however, from the application of a standard method of organization and mark procedure. This will permit quantification of hitherto purely subjective factors, will assist in the judgmental process, and permit later analysis of reliability and validity. Work in this direction is reported by the Psychological Research Wing of the Defence Science Organization, Ministry of Defence (1953). The analogous tests do not require cultural modification: the only requisite is that the tasks chosen be pertinent to jobs successful candidates will undertake. Assessment of performance on these tasks will also benefit from the use of a mark procedure, which can be standardized and validated. Fundamentally, the successful use of paper and pencil tests and selection boards depends on careful analysis of the job and careful construction of tests for maximum prediction.

The applications of the selection procedures and suggested modifications have been described above. These techniques, in a modified form, are currently being studied in actual job selection in the Indian Statistical Institute. For stenographer applicants, a battery of aptitude and performance tests has been combined with a standardized interview procedure permitting quantitative assessments. For more technical positions, special examinations are administered along with standardized interviews and group discussion. Studies reporting the use of both tests and selection boards in the Indian Services have been published in *Sankhyā* (Psychological Research Wing, 1953; Sharma, 1953).

### CONCLUSIONS

Both types of selection procedure currently used in Great Britain have applicability in India. The paper and pencil test battery is recommended for economy, and for non-administrative personnel selection. The selection board is recommended where policy-making, leadership, personal contact and group cooperation are important. Starting with careful job analysis, these techniques may be modified to achieve greater cultural meaning and to provide more standard quantitative assessments.

### INTERVIEWS

The following leading psychologists in Great Britain were interviewed in connection with this survey: Professor John Cohen, Manchester University; Professor Lee J. Cronbach, Scientific Liaison Officer, United States Navy; Mr. David C. Duncan, National Institute of Industrial Psychology; Professor Hans J. Eysenck, Institute of Psychiatry, University of London; Mr. W. D. Furneaux, Nuffield Research Unit, Institute of Psychiatry, University of London; Mr. John D. Handyside, National Institute of Industrial Psychology; Mr. K. A. G. Murray, Chairman, Civil Service Selection Board; Mr. A. K. Rice, Tavistock Institute of Human Relations; Mr. Alec Rodger, Editor of "Occupational Psychology", Birkbeck College; Professor Roger W. Russell, University College, London; Miss Margaret S. Stevenson, Research Psychologist, Civil Service Commission; and Professor Philip E. Vernon, Institute of Education, University of London.



REFERENCES

- ANASTASI, ANNE (1954): *Psychological Testing*, Macmillan, New York.
- ARKIN, H. and COLTON, R. R. (1950): *Tables for Statisticians*, Barnes and Noble, New York.
- BENNETT, G. K. (1948): *Test of Mechanical Comprehension*, Psychological Corporation, New York.
- CASTLE, P. F. C. and GARFORTH, F. I. DE LA P. (1951): Selection, training and status of supervisors :  
I. Selection. *Occup. Psychol.*, **25**, 109-123.
- CIVIL SERVICE COMMISSIONERS (1951): *Memorandum by the Civil Service Commissioners on the use of  
the Civil Service Selection Board in the Reconstruction Competitions*, H. M. Stationery Office, London.
- FRASER, J. M. (1950): New-type selection boards: a further communication. *Occup. Psychol.*, **24**,  
40-47.
- GULLIKSEN, H. (1950): *Theory of Mental Testing*, John Wiley, New York.
- HANDYSIDE, J. D. and DUNCAN, D. C. (1954): Four years later: a follow up of an experiment in selecting  
supervisors. *Occup. Psychol.*, **28**, 9-23.
- HIGHAM, T. M. (1952): Some recent work with group selection techniques. *Occup. Psychol.*, **26**, 169-175.
- PORTEUS, W. S. (1950): Psychological procedures in the selection of factory operatives. *Occup. Psychol.*,  
**24**, 113-119.
- PSYCHOLOGICAL RESEARCH WING (1953): Multiple factor analysis of personality ratings in services  
selection boards. *Sankhyā*, **13**, 17-26.
- RAVEN, J. C. (1938): *Progressive matrices Sets A, B, C, D and E.*, H. K. Lewis, London.
- SHARMA, O. C. (1953): Factor analysis of technical trades and educational examination marks of the  
Aircraftsmen of the Indian Air Force. *Sankhyā*, **13**, 27-34.
- SHIPLEY, W. C. and BURLINGAME, C. C. (1941): A convenient self-administering scale for measuring  
intellectual impairment in psychotics. *Amer. J. Psychiat.*, **97**, 1313-1325.
- STEVENSON, MARGARET S. (1951): The place of job analysis in personnel selection with special reference  
to the selection of civil servants. Research Unit Note 279, *Civil Service Commission*, London.
- VERNON, P. E. (1950): The validation of the civil service selection board procedures. *Occup. Psychol.*,  
**24**, 75-95.
- VERNON, P. E. and PARRY, J. B. (1949): *Personnel Selection in the British Forces*, University of London  
Press, London.
- WILSON, N. A. B. (1948): The work of the civil service selection board. *Occup. Psychol.*, **22**, 204-212.

*Paper received : June, 1956.*



# ISOLATION OF SOME MORALE DIMENSIONS BY FACTOR ANALYSIS

By H. C. GANGULI

*Indian Institute of Technology, Khargpur, India*

## 1. THE PROBLEM

During the war and the post-war period industrial psychology has come to occupy itself more and more with the gratifications the workers derive from the employment relationship. Many studies on problems of worker motivation and morale have been made, mostly in Western countries and more recently in India also.

1.1. As a result of some recent studies in industrial morale the 'global' or blanket concept of morale has been found to be untenable. The hypothesis that worker morale is not a single unified entity but a multidimensional concept is more and more coming to the forefront. The Survey Research Centre at Ann Arbor, which has done pioneering work in this line, has isolated five distinct morale dimensions or types of satisfactions that the individual derives from the industrial situation (Katz, 1950). These are intrinsic job satisfactions, satisfactions derived from involvement in the immediate work group, from identification with the company, from interpersonal relations with the supervisor as a personality and the indirect satisfactions of the individual's needs from his membership of the organisation. It may be noted that the term morale is given here an individual-organic emphasis and is used synonymously with the satisfaction of the worker from the employment relationship.

1.2. In India, so far as the author knows, no attempt has been made to isolate and identify the different morale dimensions. Also, although the attitude survey method is mostly used for obtaining measures of worker morale, it is not known how the different items in an attitude scale are related to each other or are loaded with these morale factors so far as the Indian worker is concerned. To know this is important since it is possible that the same or similarly worded statements may be tapping different attitudes in different social groups.

## 2. THE STUDY

2.1. A morale survey was made by the author in an important electric fan manufacturing concern of Calcutta. This factory produces more than 29% of total fans manufactured in India and had 1,890 workers on its roll in 1951. The survey involved all the 550 workers in the foundry shop, the machine shop and one assembly shop. The method of study closely resembled the sample interview survey technique (Cartwright, 1950, Likert, 1947), though on a small scale. The study was conducted on the basis of an attitude scale constructed on the principle of summated ratings (Likert, 1932). This scale had 41 items, 32 of these were 5 point statements and 9, 3 point statements. It was constructed on the basis of initial exploratory interviews with 40 workers and four pretests. This scale contains all the items that were found to be significant and meaningful for these workers and nothing besides. The scale is of such a nature that with minor modifications it can be appropriately used on other workers in the engineering industry.

2.2. The statements had fixed alternative answers and the relevant response alternatives were checked by the investigator during an open non-directional interview with the worker. The usual precautions in the framing of these items, and conduct of the interviews for ensuring the validity of results were taken (Anonymous, 1949; Marriott, 1953; Oldfield, 1947; Vernon, 1952).

## 3. THE FACTOR ANALYSIS

3.1. Hotelling and Kelly's method of principal components and Thurstone's centroid method represent the two major methods of factor analysis. Thurstone's centroid solution has the advantage of making no assumption regarding the independence of the factors and also yields more consistent factors. So this method has been adopted for analysing the attitudinal data obtained from the above survey.

3.2. The factor analysis was conducted with only 23 of the 41 items as otherwise the task would have been very laborious. The figure 23 is arbitrary. These 23 items were selected for their 'goodness', as judged from their discrimination values or *d*-values. The *d*-values were calculated from two criterion groups made up of 50 most satisfied or high morale workers and 50 least satisfied or low morale workers in a larger group of 175 workers, the satisfaction or dissatisfaction of the worker being indicated by his score on the full preliminary scale of 41 items. *D*-values for 38 of these items were positive and significant. 17 of the 23 items selected for factor analysis had the highest *d*-values. The other six items have been included not on the strength of their *d*-values (which however were positive and significant) but in order to give representation to some specific aspects, viz., attitude towards the job content and satisfaction with pay increases and the welfare activities of the company.

3.3. Table 1 below, gives these 23 attitudinal items. Table 2 gives the correlation matrix of the 23 items based on data from 175 workers.

TABLE 1. ATTITUDINAL ITEMS INCLUDED IN FACTOR ANALYSIS

- 
1. there are some other work I would be able to do better than the work I do now.
  2. like the job on the whole.
  3. the company's policy is to overdrive the worker and get the maximum out of him.
  4. satisfied with present earnings.
  5. income is somewhat larger than what I would have got in other similar factories.
  6. satisfied with the method of allocation of increments.
  7. satisfied with chances of increasing income at this factory.
  8. shall not lose job so long as I work efficiently.
  9. satisfied with my chances of getting a better type of job.
  10. immediate superior is reasonable in the work he expects from me.
  11. he gives reasonable attention to suggestions regarding method of work, tools etc.
  12. he is my own man.
  13. satisfied with the general supervision of my section.
  14. satisfied with the allotment of work in the section.
  15. my section in-charge is good in the way he handles the workers.
  16. I am given maximum facilities for doing my work properly.
  17. satisfied with the running of the factory canteen.
  18. factory dispensary gives satisfactory service.
  19. the existing leave rules cover my average requirements adequately.
  20. I feel that this company treats its workers worse than other companies.
  21. the company is sympathetic to and appreciative of the worker's point of view.
  22. this factory is a better place to work than other neighbouring factories.
  23. rarely think of quitting this company.
- 

3.4. The factor analysis of the above data has been largely based on computational procedures given by Guilford (1936).

## ISOLATION OF SOME MORALE DIMENSIONS BY FACTOR ANALYSIS

TABLE 2. INTERCORRELATION OF THE 23 ITEMS (N = 175)

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)
1	.35	.13	.15	.05	.01	.05	-.07	.05	-.15	.02	.03	-.01	.17	-.06	-.30	-.03	.04	.21	.17	.09	.01	-.29
2		.15	.26	.18	.19	.21	-.06	.14	.11	.20	-.14	.09	.29	-.02	.19	.03	.07	.13	.11	.13	.34	-.35
3			.48	.59	.36	.43	.38	.30	.22	.10	.17	.28	.28	.14	.15	.40	.29	.36	.49	.65	.41	.39
4				.56	.43	.32	.20	.34	.17	.21	.10	.28	.29	.23	.28	.68	.25	.24	.36	.49	.32	.23
5					.30	.84	.27	.23	.19	.21	.07	.25	.33	.18	-.18	.34	.35	.21	.40	.55	.37	.38
6						.54	.13	.39	.15	.06	.09	.27	.32	.19	.22	.27	.27	.18	.23	.01	.31	.18
7							.24	.52	.27	.08	.24	.34	.30	.29	-.14	.30	.44	.38	.34	.44	.44	.38
8								.29	.25	.22	.31	.39	.32	-.20	.06	.27	.30	-.32	.43	-.18	.28	.24
9									.20	.19	.30	.43	.40	.38	.52	.30	.38	.15	.41	.44	.31	.08
10									.66	.53	.53	.32	.28	.27	.10	.08	.30	-.50	.21	.21	.18	.15
11										.58	.41	.25	.25	.90	.07	.18	.21	.07	.32	.06	.24	-.31
12											.29	.09	.30	.11	.01	.01	.15	.06	.25	.15	.09	.39
13												.53	.74	.38	.34	.37	.20	.42	.39	.38	.25	
14													.76	.58	.24	.18	.13	.42	.34	.37	.28	
15														.45	.01	.17	.04	.36	.23	.22	.16	
16																-.12	.10	-.06	.21	.21	.17	-.01
17																	.44	.38	.37	.23	.27	.08
18																		.45	.39	.41	.31	.18
19																			.42	.38	.32	.42
20																				.64	.44	.19
21																					.42	.30
22																						.30



3.5. The following points may be mentioned regarding this analysis.

i) The communality of each test item has been taken as equal to the highest correlation in the column from that item.

ii) Reflection had to be resorted to in the case of some items of every factor residual for maximising positive signs.

iii) To know the point where enough factors have been taken out and further analysis should be discontinued, an empirical criterion developed by Tucker has been employed. The rule is given by the relation,

$$\phi = \sqrt{\frac{\Sigma \rho_{s+1}}{\Sigma \rho_s}} = \frac{n-1}{n}$$

where  $n$  is the number of test variables in the correlation matrix,  $\Sigma \rho_s$  is the sum of the absolute values of the residuals after  $s$  factors and with adjusted diagonals and  $\Sigma \rho_{s+1}$  is the sum of the absolute values of the residuals after  $(s+1)$  factors with original or unadjusted diagonal values.

If  $\phi$  equals or exceeds  $\frac{n-1}{n}$  then  $S$  is really the last factor needing extraction.

In this analysis, this test was applied after the 4th factor loadings were calculated and 4th factor residuals computed. To know whether the 3rd is really the last factor the following statistics were calculated.

$$\Sigma \rho_3 = 50.722 ;$$

$$\therefore \phi = 0.9624$$

$$\Sigma \rho_4 = 46.984 ;$$

$$\frac{n-1}{n} = 0.9565$$

Since  $\phi$  exceeds  $\frac{n-1}{n}$ , it may be accepted that the 3rd factor is the last factor that should be extracted from this correlation matrix.

iv) Table 3 gives the original factor loadings for the different test items.

TABLE 3. FACTOR LOADINGS OF THE 23 TEST ITEMS ON THE THREE CENTROIDS PRIOR TO ROTATION OF AXES

items	factor loadings			items	factor loadings		
	I	II	III		I	II	III
1	.083	.207	-.238	2	.252	.127	-.407
3	.667	.309	.202	4	.654	.248	-.122
5	.642	.407	.348	6	.482	.110	-.122
7	.692	.396	.270	8	.357	-.156	-.293
9	.622	-.151	.071	10	.416	-.439	-.427
11	.498	-.548	.089	12	.406	-.442	-.380
13	.691	-.339	.078	14	.685	-.213	.284
15	.568	-.582	.391	16	.305	-.421	.279
17	.492	.265	-.109	18	.556	.126	.086
19	.368	.621	-.308	20	.703	.058	-.018
21	.619	.310	.054	22	.593	.152	-.048
23	.345	.129	.495				

## ISOLATION OF SOME MORALE DIMENSIONS BY FACTOR ANALYSIS

3.6. Thurstone believes that if the factors should have psychological as well as mathematical meaning, and would remain invariant even if the test was analysed as part of a different battery, the centroid factors must be rotated. This is necessary not only for the first centroid which includes almost of everything that is in the battery but also for the second and subsequent factors which have positive weights on some test items and negative weights on others. In this study although only three factors have been extracted, the rotations have been done on the more general method of taking only two factors at a time, rotating them about the third axis and computing the new factor loadings after this rotation. This procedure has been repeated as many times as necessary. The location of the new axis has been guided by the principle of maximisation of zero factor loadings and minimisation of negative loadings.

3.7. Altogether three rotations had to be made. The first rotation was about Axis I, in the plane II--III, the angle of rotation being  $49^\circ$  (clockwise). The second rotation of  $66^\circ$  was of factors I and II' around the axis III'. This was followed by a third rotation of  $58^\circ$  in the plane II''--III' about the axis I'. The final loadings of the test items on the three factors indicated by the symbols  $C$ ,  $Sp$  and  $S_0$  are given below in Table 4.

TABLE 4. FINAL FACTOR LOADINGS OF 23 TEST ITEMS ON THE THREE COMMON FACTORS

test item	$C$	$Sp$	$S_0$	communality	uniqueness
1	.173	.108	-.254	.1061	.8939
2	.239	.351	-.254	.2448	.7552
3	.728	.024	.225	.5812	.4188
4	.651	.280	.033	.5033	.4967
5	.782	-.142	.257	.6977	.3023
6	.432	.267	.046	.2600	.7400
7	.751	-.012	.284	.6448	.3552
8	.153	.461	.036	.2372	.7628
9	.393	.326	.393	.4152	.5848
10	.010	.727	.138	.5477	.4523
11	.049	.451	.593	.5575	.4425
12	.005	.690	.168	.5043	.4957
13	.329	.448	.538	.5984	.4016
14	.422	.234	.602	.5952	.4048
15	.373	.449	.152	.3638	.6362
16	-.003	.157	.569	.3484	.6516
17	.538	.186	-.034	.3252	.6748
18	.517	.144	.210	.3321	.6679
19	.647	.090	-.435	.6159	.3841
20	.579	.324	.238	.4969	.5031
21	.678	.106	.104	.4817	.5183
22	.551	.245	.117	.3773	.6227
23	.391	-.254	.404	.3806	.6194

3.8. To determine the uniqueness of any factor solution, Thurstone has put forth his theory of 'simple structure'. He has laid down three empirical conditions to determine how far the factor loadings satisfy the test of 'simple structure'. These are :

- (a) at least one zero loading in each row;
- (b) at least as many zero loadings in each column as there are columns;

- (c) at least as many  $XO$  or  $OX$  entries in each pair of columns as there are columns; by an  $XO$  entry is meant a loading in each column opposite a zero in the other. (Thomson, 1950).

3.9. As regards the loadings of the 23 items on the three factors arrived at after rotation (Table 4) it will be seen that conditions (b) and (c) are satisfied. But the first condition is not fully satisfied. Only 11 rows out of 23 have one zero loading each. This means that although no factor is general and the test items are qualitatively different, there are nevertheless some items in which all the three factors are present. This defect has arisen in spite of all precautions taken at the time of construction of the preliminary scale to ensure that one item refers only to one single variable or factor. How far construction of attitude items is possible such that to a large group of persons it will always have the same connotation and response to which will be dictated by reference to the same issue is a debatable question. Thus although the factor loadings of this scale do not fully satisfy Thurstone's conditions of a 'simple structure' (which even a test of mental ability seldom does) its findings are nevertheless of much practical value and fully in accord with the logical analysis of the field.

#### 4. IDENTIFICATION OF THE PRIMARY FACTORS

4.1. "The findings of factors in an analysis simply indicates the presence of some common causes or determiners in the variables analyzed" (Wolfe, 1940) and almost the only way of identifying them is to examine which tests or test items are most highly saturated with them; in other words, on the basis of content analysis of the relevant tests. Thus the first factor, for example, in Table 4 is present to a substantial degree (accounting for more than 10% of the total variance of the item) in 16 items. These 16 items refer to the worker's satisfaction with wages, with supervision, and company policies and practices. It finds best expression in items referring to satisfaction with wages and the chances a person has got in increasing his income. It also runs through items referring to other aspects of the organisation, like the effectiveness of supervision, the company's overall personnel policy, its welfare activities etc. This factor, it seems, combines within itself aspects of what Katz (1950) has called identification with the company, i.e., satisfactions derived from membership of the total organization and the indirect satisfactions that he derives because of the needs that he can satisfy outside the company through the wages he gets etc. It also seems that underlying this factor and very closely associated with it is the feeling of confidence the worker has in the company, when by confidence is meant "the genuine emotional conviction that one will be able reasonably to satisfy one's needs through this employment relationship" (McGregor, 1949). This factor has therefore been referred to as factor  $C$ .

4.2. The second and third factors in Table 4, called factors  $S_p$  and  $S_o$  respectively, are easier to interpret. Items carrying significant loads of these factors (of more than 0.32) refer directly or indirectly to the supervisor and the nature of supervision. A further examination of the contents of these items shows some clear-cut differences between the two factors. Items highly saturated with factor  $S_p$  refer to the reasonableness of the first-line supervisor in the work he expects from the worker, the attention he gives to his various suggestions the extent to which the worker can regard him as 'his own man', the skill with which the foremen handles his workers etc. Factor  $S_o$ , on the other hand, is covered by items having to do with facilities given to the worker for doing his job properly, allotment



and division of the work-load, the way in which his suggestions for improving methods of work, tools etc., are received and so on: in short, it has reference to the technical and organisational efficiency of the shop supervision. Factor  $S_p$  thus seems to refer to the human and personal aspects of supervision whereas factor  $S_0$  refers to the organisational and technical aspects of supervision. That workers make a distinction between these two aspects of supervision is now very much recognised.

4.3. It may be noted that a worker's liking for his job and his feeling of security at it seem to be largely determined by or related to his personal relations with his superiors. (factor  $S_p$ ) It is again this relation he has with his boss and the treatment he receives from him that serve as an indicator to him of the way in which the company tries to treat its workers. If the boss is good, the company is good and conversely also. On the other hand, it may also be noted that the frequency with which a person thinks of quitting his job seems to be closely related to the shop organisation and the efficiency with which it is run (factor  $S_0$ ).

4.4. The above interpretation of the nature of the three factors seems to the present investigator as essentially logical. But as has been pointed out by others, the name given to a factor is merely the experimenter's hypothesis regarding the nature of that factor. Only further studies can show how far the present results are generally applicable, for these are affected to some extent by the nature of the scale, the characteristics, personal and otherwise, of the subjects, by the way of scoring of the scale etc.

## 5. SUMMARY AND CONCLUSION

5.1. Recent studies have shown that the unitary concept of morale is not adequate. Rather, morale is a multi-dimensional concept and can be broken down into a number of factors. To determine what these factors are and how they saturate the different items in an attitude scale, a factor analysis was done with data collected from a morale survey of workers in a Calcutta engineering factory. To ensure that items making up the scale used for measuring morale are valid and relevant to the workers concerned, the scale was constructed after extensive preliminary interviews and four pretests.

5.2. The results of the factor analysis have revealed the presence of three factors or dimension of morale. The first of these, called factor  $C$ , refers to the worker's satisfaction with the total organisation as well as with the benefits he derives from it. It refers to issues like how the factory compares with other factories as a place to work, whether the company is generally sympathetic to the workers and appreciative of their point of view, whether the income is satisfactory etc. In short, this factor refers to the feeling of confidence the individual has got in the company and the conviction he has of being able to satisfy his needs from this employment relationship.

5.3. The other two factors refer to satisfaction with supervision. The men distinguish between satisfaction derived from inter-personal relations with the supervisor as a person and satisfaction with his technical competence and running of the shop. The first factor,  $S_p$ , has reference to such items as the supervisor's skill in handling the men, his reasonableness in what he expects from them and in general, how far he can be regarded as their 'own men.' Factor  $S_0$  has reference to the effectiveness and competence with which the supervisor performs his duties like division of work-load, attention to workers'

suggestions regarding methods of work etc., giving facilities to his men for doing the work properly and so on.

5.4. The most conspicuous among those items which have high specificity is the one that refers to how far the worker is satisfied as to his ability to do the job properly and if there is some other job which he may be able to do better. It is expected that this item may have reference to another morale dimension bound up with the nature of the job itself. Such a factor as intrinsic job satisfaction, that is, satisfaction with the nature of work itself has been isolated by other investigators.

5.5. In conclusion it may be said that the three factors isolated above together with the one referring to satisfaction with the job content do not exhaust all the possible morale dimensions. Other factors may be there. Satisfactions derived from membership of the immediate work group, from being part of an occupational system etc., have been mentioned by others. Further investigations are necessary to know how far relevant these factors are for the Indian workers and how do they load the different items. There is no reason to believe however that sources of satisfaction for the Indian worker is very much different from those of workers in other countries.

#### REFERENCES

- ANONYMOUS (1949): *Human Relations Study Technique*, Survey Research Center, Institute for Social Research, Ann Arbor.
- CARTWRIGHT, D. (1950): Survey Research: Psychological Economics; Chapter 4, *Experiments in Social Process*. (Edited by J. G. Miller), McGraw Hill Book Co., New York.
- GUILFORD, J. P. (1936): *Psychometric Methods*, McGraw Hill Book Co., New York.
- KATZ, D. (1950): An over view of human relations programme. Human Relations Programme of the Survey Research Center, Ann Arbor.
- LIKERT, R. (1932): A technique for the measurement of attitudes. *Archives of Psychology*, 140, New York.
- , (1947): The sample interview survey. *Current Trends in Psychology*, (Edited by W. Dennis) 223-225, University Pittsburgh Press.
- MARRIOTT, R. (1953): Some problems in attitude survey methodology. *Occupational Psychology*, July.
- MCGREGOR, DOUGLAS, D. (1949): Toward a theory of organised human effort in industry. Part III. *Psychology of Labour-management Relations*. (Edited by Kornhauser, A.), Industrial Relations Research Association, U.S.A.
- OLDFIELD, R. C. (1947): *The Psychology of Interview*, Methuen and Co., London.
- THOMSON, G. (1950): *The Factorial Analysis of Human Ability*. University of London Press, London.
- VERNON, P. E. (1952): *The Assessment of Psychological Qualities by Verbal Methods*. Industrial Health Research Board, Report No. 83, London.
- WOLFLE, D. (1940): *Factor Analysis to 1940*. University of Chicago Press, Chicago.

Paper received : October 1955

# INVERSION OF $25 \times 25$ MATRIX ON 602A CALCULATING PUNCH

By DEB KUMAR BOSE AND AMAL KUMAR ROY

*Indian Statistical Institute, Calcutta*

## INTRODUCTION

Systems of linear equations in several variables appear in analytical work relating to almost all branches of sciences. For their solution the inversion of a given matrix of coefficients of the unknowns is often found useful. Punched card machines have been used variously for the solution of such equations especially when there are large number of unknowns. A problem requiring the inversion of a matrix of order  $25 \times 25$  recently arose in the course of certain econometric studies conducted in the Indian Statistical Institute under the guidance of Ragnar Frisch in connection with Planning for National Development. This paper deals with the punched card method adopted for the inversion of them atrix. F. M. Verzuh (1949) indicated a method of solution of simultaneous equations, with the aid of the 602 Calculating Punch, dealing with matrices of the order  $10 \times 10$ . We have adopted here a different means for the solution by inverting a matrix of order  $25 \times 25$  through methods involving simpler machine operations and substantial saving in machine hours, although the plugging required has become little complicated.

## 1. MATRIX INVERSION

The process of matrix inversion adopted is indicated below. Details are given by Rao (1952).

Let the matrix to be inverted be,

$$A = \begin{matrix} a_{11} & \dots & a_{1n} \\ a_{21} & \dots & a_{2n} \\ \dots & \dots & \dots \\ a_{n1} & \dots & a_{nn} \end{matrix}$$

(a) Divide row 1 by  $a_{11}$ , multiply the new row 1 by  $a_{i1}$  and subtract from this product row  $i$  for  $i = 2, 3, 4, \dots, n$ , successively. This set of operations result in 1 in the first position, and zeros elsewhere in the first column of the matrix. Again applying the same method on row 2, we get 1 in the second position of the main diagonal and zeroes everywhere else in the second column. By repeating this process on the subsequent rows, the given matrix is reduced to a triangular matrix of the following form having 1 in each of its diagonal elements.

$$\begin{matrix} 1 & * & * & \dots & * \\ 0 & 1 & * & \dots & * \\ 0 & 0 & 1 & \dots & * \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \end{matrix}$$



(b) Now take this triangular matrix and multiply row  $n$  by the  $n$ -th element of row  $i$  and subtract it from the  $i$ -th row where  $i = 1, 2, \dots, n-1$ , this operation results in all zeros in the  $n$ -th column of row  $1, 2, \dots, n-1$ . Again, applying this method on row  $n-1$  we get all zeroes in the  $(n-1)$ th column of row  $1, 2, \dots, n-2$ . Repeat this process on row  $n-2, \dots, 2, 1$ , until this triangular matrix is reduced to a unit matrix  $I$ .

Simultaneous application of the same sequence of operations on a unit matrix, when taken on the right-hand side along with the matrix  $A$ , leads to the transformation of the unit matrix into  $A^{-1}$ , the inverse of  $A$ .

The operational steps may be summarised here. First we take the matrix  $A$  and a unit matrix  $I$ :

$$[A|I]$$

Performing the operation (a) above, we have,

$$[B|C]$$

where  $B$  is a triangular matrix having 1 in each of its diagonal elements.

Then performing operation (b), we have,

$$[I|A^{-1}]$$

## 2. MACHINE OPERATIONS

The detailed steps are stated below.

(1) Individual cards are punched for each of the elements of the matrices  $A$  and  $I$  with respective row and column numbers for identification according to the following punch card design.

DESIGN OF CARD			
sl. no.	item	card columns	remarks
1.	step no.	1-2	
2.	row no. ( $i$ )	3-4	
3.	col. no. ( $j$ )	5-6	
4.	value of the element ( $a_{ij}$ )	7-15	(algebraic sign in col. 7: 0 for positive, 5 for negative; 6 places of decimals)
5.	either quotient $\left( B = \frac{a_{ij}}{a_{ii}} \right)$		
6.	or product difference ( $\pm S$ ) designation	16-24 80	(sign in col. 16, 6 places of decimals) (X for $P$ cards, 1 for $Q$ cards, blank for $R$ cards)

## INVERSION OF $25 \times 25$ MATRIX ON 602A CALCULATING PUNCH

- (2) The cards containing the elements of the matrices  $A$  and  $I$  are sorted according to row and column numbers (i.e.,  $i$  and  $j$ ).
- (3) Separate the first row of the matrix (i.e. cards with  $i = 1$ ).
- (4) Remove the first card (i.e.,  $a_{11}$ ) and punch  $X$  on it in col., 1.
- (5) Divide the elements  $a_{ij}$  on the other cards of the first row by  $a_{11}$  on the leading column card (card with  $X$  in col. 1). Punch the quotients on the dividend cards.
- (6) Reproduce the dividend cards punching the quotients in the original field for elements. These cards will form the pivotal row. Punch 1 in col. 80 on these set of cards (call them  $Q$  cards).
- (7) Take the cards of the next row of the matrices. Punch  $X$  on col. 80 of the first card (call it  $P$  card and the remaining cards of the row  $R$  cards).
- (8) Place the  $Q$  cards followed by the  $P$  and  $R$  cards and sort them on cols., 5-6 (i.e., for  $j$ ).
- (9) Pass the cards through 602A to calculate  $(\pm P) - (\pm Q_j) - (\pm R_j) = \pm S_j$ , ( $\pm S_j$  being punched on the  $R_j$  cards). Details of this step are given in Section 3.
- (10) Separate the  $Q$  cards from the  $P$  and  $R$  cards by sorting on col., 80.
- (11) Reproduce the  $R_j$  cards to bring back  $S_j$  in the original field for elements.
- (12) Perform sum-check on Tabulator.
- (13) Take the cards of the next row of the composite matrix and repeat operations from 8 onwards till the last row is exhausted. Then the pivotal row ( $Q$  cards) is eliminated.
- (14) Take all reproduced  $R$  cards (i.e., new  $S$  cards) which forms a new matrix with the number of rows reduced by one. Repeat steps (3) to (14) till all the pivotal rows are eliminated.
- (15) Take all the pivotal row cards ( $Q$  cards)—eliminated by earlier operations. Sort them for  $i$  and  $j$ . They will now form new matrices  $B$  followed by  $C$ . The rectangular matrix constituted by the last column of the matrix  $B$  and the entire matrix  $C$  is separated out.
- (16) Take the last row of the matrix so formed and treat it as the pivotal row. Punch 1 in col. 80 on this set of cards. Call them  $Q$  cards as in (6). Repeat procedures (7) to (13) with one difference at the calculating stage (9) where the formula will be  $-[(\pm P)(\pm Q_j) - (\pm R_j)] = \pm S_j$ .
- (17) Take all reproduced  $R$  cards (i.e. new  $S$  cards) which forms a new matrix with the number of rows reduced by one, as before. At this stage, insert the next to last column vector from the  $B$  matrix in the matrix at hand, Then repeat step (16) till all the columns of the  $B$  matrix are exhausted.

The matrix formed by these  $Q$  cards will represent the required matrix  $A^{-1}$ .

### 3. COMPUTATION ON 602A CALCULATING PUNCH

The calculation work would consist of two types of operations in the 602A calculating punch, the first involving the division  $\frac{a_{ij}}{a_{ii}}$  for building up of the pivotal row and the second

giving rise to the expression :  $(\pm P) \cdot (\pm Q) - (\pm R) = \pm S$ . The breaking up of the machine operations into two distinct parts would prove very advantageous in the setting up of the machine. The first operation is required 25 times for the building up of 25 pivotal rows. The second type of operations would continue for the rest of the calculations in two stages as indicated above.

The panel wiring required is shown in the diagram, where it may be observed that factor  $P$  is indicated by  $X$  in col. 80.

"	$Q$	"	1	"
"	$R$	"	blank	"

0 is punched in col. 7 for positive sign and 5 punched for negative sign (so that  $5+5=0$ ,  $0+5=5$ ) as suggested by Matthai (1950).

*Read cycle :* 2 pilot selectors (nos. 1 and 2) are picked up from the control brush set on card column 80, by using  $X$  and 1 in card column 80.

$P$	is read in storage $1R$ , its sign in storage $1L$
$Q$	" 2 " counter 4
$R$	" 4 " storage $3R$

Cards for  $P$  and  $Q$  are read and skipped off. Cards for  $R$  are held at the punch bed till the calculations are completed and the result  $S$  punched on it.

*Programme 1 :* Multiplicand  $Q$  is read out of storage  $2R$  and read in counters  $1+2+3$  to develop the product  $PQ$ .

*Programme 2 :*  $PQ$  is transferred from counters  $1+2+3$  to counters  $5+6$  corrected to 6 places of decimals. The sign for  $PQ$  is developed in counter 4 by adding digit in storage  $1L$  to the digit in counter 4.

*Programme 3 :* Three pilot selectors (3, 4 and 6) energized by readings of digit 5 in counter 4 and storage  $3R$ . Two of the Pilot selectors are picked up by the digit in counter 4 and the third one by that of storage  $3R$ .

*Programme 4 :*  $R$  from storage  $4L+4R$  is read into counters  $5+6$  in true form when the pair of selectors 4 and 5 are unequal (i.e. only one of them is picked up) and in complementary form when selectors are equal (both picked up or dropped).

*Programme 5 :* Co-selector number 5 is picked up through pilot selector 5 energized by complementary figures in counters  $5+6$ . Digit 5 from emitter is taken to the transfer side of co-selector 5 through Pilot selector 6 which moves by digit 5 in counter 4. Digit from counter 4 is read out and put directly to  $N$  hub of the co-selector.  $S$  is read out of counters  $5+6$  and put in storage 6 for punching, while its sign is furnished to Punch Storage 6 through the co-selector 5. The card for  $R$  is skipped off after  $S$  is punched on it.

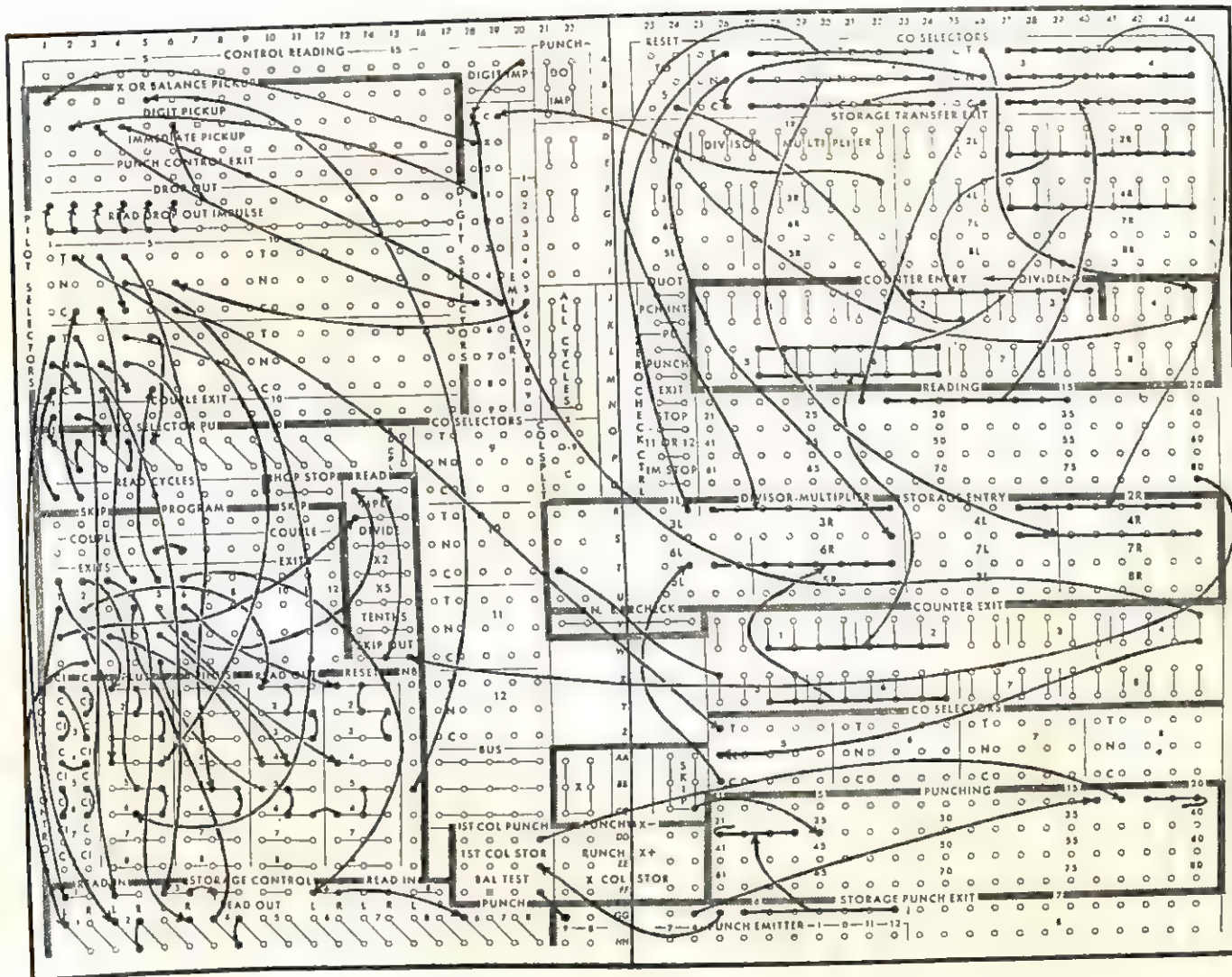
It will be seen that in the equation  $(\pm P) \cdot (\pm Q) - (\pm R) = (\pm S)$ . The sign of  $S$  opposite to that of the product  $PQ$  (as registered in counter 4 at Programme 2) when  $PQ$  and  $R$  are of same signs and  $R$  is greater than  $PQ$  (indicated by complementary figures



# INVERSION OF $25 \times 25$ MATRIX ON 602A CALCULATING PUNCH

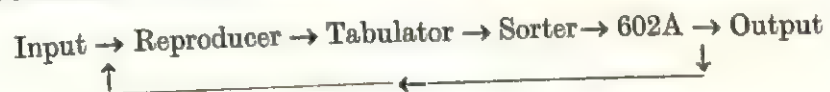
in counters 5+6). In all the other cases, the sign of  $S$  is same as that of  $PQ$ . This is effected by the wiring shown in the diagram.

WIRING DIAGRAM



## 4. MERITS OF THE METHOD

About 75 calculating machine hours are required to be spent for working out the  $25 \times 25$  matrix. The merit of the procedure detailed above lies in the fact that machine operations are reduced to a simple circular flow of the following form:



In contrast to Verzuh's method of punching two elements in one card, individual cards are punched for each of the elements, thus keeping the factors  $P$  and  $Q$  dissociated. Machine

operations are made simpler resulting in a substantial saving of machine hours, while complication if any is introduced only in the wiring.

We are grateful to Shri A. Halder for valuable suggestions, and to Shri S. Dutta, for participation in discussions, in the course of preparation of the paper.

#### REFERENCES

- RAO, C. R. (1952): *Advanced Statistical Methods in Biometric Research*, 30-31, John Wiley & Sons, New York.
- VERZUH, FRANK M. (1949): The solution of simultaneous equations with the aid of 602 calculating punch. *Mathematical Tables and Aids to Computations*, July 1949, 453-55.
- MATTHAI, A. (1950): On methods of handling algebraic signs on the Hollerith multiplier. *Sankhyā*, 10, 124-128.

*Paper received : June, 1955.*





















